EYE TRACKING METHODS FOR ANALYSIS OF VISUO-COGNITIVE BEHAVIOR IN MEDICAL IMAGING

A Dissertation

by

FOLAMI TOLULOPE ALAMUDUN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Tracy Anne Hammond
Committee Members,	Thomas R. Ioerger
	Tiffani L. Williams
	Thomas Ferris
Head of Department,	Dilma Da Silva

May 2016

Major Subject: Computer Science

ABSTRACT

Predictive modeling of human visual search behavior and the underlying metacognitive processes is now possible thanks to significant advances in bio-sensing device technology and machine intelligence. Eye tracking bio-sensors, for example, can measure psycho-physiological response through change events in configuration of the human eye. These events include positional changes such as visual fixation, saccadic movements, and scanpath, and non-positional changes such as blinks and pupil dilation and constriction. Using data from eye-tracking sensors, we can model human perception, cognitive processes, and responses to external stimuli.

In this study, we investigated the visuo-cognitive behavior of clinicians during the diagnostic decision process for breast cancer screening under clinically equivalent experimental conditions involving multiple monitors and breast projection views. Using a head-mounted eye tracking device and a customized user interface, we recorded eye change events and diagnostic decisions from 10 clinicians (three breast-imaging radiologists and seven Radiology residents) for a corpus of 100 screening mammograms (comprising cases of varied pathology and breast parenchyma density).

We proposed novel features and gaze analysis techniques, which help to encode discriminative pattern changes in positional and non-positional measures of eye events. These changes were shown to correlate with individual image readers' identity and experience level, mammographic case pathology and breast parenchyma density, and diagnostic decision.

Furthermore, our results suggest that a combination of machine intelligence and bio-sensing modalities can provide adequate predictive capability for the characterization of a mammographic case and image readers diagnostic performance. Lastly,

features characterizing eye movements can be utilized for biometric identification purposes. These findings are impactful in real-time performance monitoring and personalized intelligent training and evaluation systems in screening mammography. Further, the developed algorithms are applicable in other application domains involving high-risk visual tasks.

DEDICATION

This work is dedicated to Oluwaseyi Alamudun, without who's unwavering support none of this could be at all possible; not even by a long chalk! And to my super clan: Abiodun, Oluwadarasimi, Adetutu, Oluwafolahanmi, and to the little one who is yet to arrive. I am humbled and thankful for your smiles, laughter, and love.

And my deepest gratitude to Him, through Who's Grace I came into being and am permitted to exist!

ACKNOWLEDGEMENTS

First, I wish to extend my deepest gratitude to my advisor, Dr. Tracy Anne Hammond for her support as an advisor, a mentor, and a guide through much of the challenges I encountered during my graduate experience. I am always in awe at your brilliance both in the manner in which your fulfill your professional duties and in your exemplary character as a human being. I thank you for believing in me, for inspiring me through your passion for both work and play, and for supporting me. I will also like to thank Dr. Georgia Tourassi, who has been a strong mentor to me. I cannot thank you enough for taking me under your wing and providing immeasurable help in propelling my research through guidance, ideas, unwavering support, and through the many opportunities to conduct collaborative research. Many thanks to Dr. Hongjoon Yoon, for being such an awesome mentor and friend. Thank you for being so welcoming and keeping me grounded in my research whenever I let my imagination get the best of me. A special thanks to Dr. Lawrence Johnson, who has been a mentor to me since my first days as an undergraduate at the University of Texas at El Paso.

I wish to thank the members of my committee: Dr. Ioerger, who has offered his unwavering support as a mentor and teacher since I first set foot on the campus of Texas A&M University. Many, many thanks to Dr. Tiffani Williams, who has mentored, counseled, and advised me through the difficult and also the not so difficult experiences during my graduate program. And to Dr. Thomas Ferris for his inspiring contributions and guidance.

I wish to thank my peers at the Sketch Recognition lab, with whom I have had nothing but pleasant and inspiring experiences. First, a special thank you to

Stephanie Valentine for being an inspiring colleague and for the memorable conversations. Thank you to Manoj Prasad, Danielle Cummings, Paul Taele, Ayden Kim, Murat Russell, JongIn Koh, Sunah Park, Vijay Rajanna, Anna Stepanova and other members. Many thanks to my former colleagues Jongyoon Choi, Sandesh Aryal, Joseph Lee, and Gene Huang. Thanks to my other peers at Texas A&M University who have helped make this journey memorable. A special thank you to the departmental staff for their assistance over the years: Karrie Bourquin, Elena Rodriguez, Sybil Popham, Valerie Sorenson, Leslie Darling, Dave Cote, Bruce Veals, Theresa Roberts, and a special thanks to Kathy Waskom and other departmental staff members. Thank you kindly.

Thank you to my friends who have stuck with me through this process and put up with my many changes: Mighty & Salewa Itauma, Segun & Erma Williams, Soji & Lanre Awe, and Charles & Kemi Daramola, Samson & Yemisi Vese, Funke Owolabi, and Olusola Ilupeju. A special thank you to my siblings, Yele, Yemi & Alaba, and Jimi & Temi Alamudun, Kemi, Sesan, Taiye, and Kehinde Owonubi. And to Elizabeth Ayorinde, Damola Alamudun, Mike & Lara Ojowa, Sola & Funmi Oyewole, Kehinde & Yinka Alamudun, My deepest gratitude to my parents, Folagbade & Oluwatoyin Alamudun, and John & Olufunke Owonubi, who have provided their unwavering support and encouragement throughout my graduate experience. “E seun mo dupe lopo lopo!”

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

1. INTRODUCTION

Diagnosis in medical practice is a broader term describing the process of gathering information about a patient, processing the acquired information, and finally aggregating this information to classify the patients condition into one of many well-defined category. The specificity of the latter category enables medical practitioners to make more meaningful medical decisions about treatment and prognosis. In short, medical diagnosis described in computational terms, is a subroutine, which involves data acquisition, data processing, and class identification.

The medical diagnostic process generally begins with a process of information gathering and information processing. Information about a patient is acquired by probing patient history and a general physical exam. With technological and computational advancements over time, medical practitioners now have access to more sophisticated techniques, such as medical imaging, for obtaining previously unavailable direct information about the internal anatomy of a patients body. More formally, medical imaging describes the process of acquiring a visual representation of the internal structures within a patients body (underneath the skin).

Medical imaging has revolutionized health care over the past several decades. Advances in imaging technology have led to the development of numerous image data acquisition modalities, such as X-ray radiography, magnetic resonance imaging (MRI), medical ultrasonography, endoscopy, elastography, tactile imaging, thermography, medical photography, computed tomography, and a large collection of nuclear based functional imaging techniques (such as positron emission tomography). These data allow for the characterization of anatomical state, metabolic processes, and other functions pertaining to body tissues, the assessment of which, aid the prac-

titioner in determining if an abnormality is present, where it is located, and other important case specific characteristics.

Imaging technology requires specialized training for data acquisition, and more importantly, for data interpretation. One area of speciality that focuses on the latter is radiology. Radiology is a branch of medicine that specializes in application of medical imaging technology in the diagnoses and treatment of injuries or diseases. Medical images are typically acquired by a radiographer (or radiologic technologist), while the image reading and interpretation is performed by a diagnostic radiologist (or radiographer).

One area of medical diagnosis where the application of medical imaging has had a significant impact is in cancer diagnosis and treatment. *Cancer*, in medicine, refers to a class of genetic diseases characterized by an uncontrolled growth and subsequent spread of abnormal cells. Cancerous cells (tumors) are caused by changes in the genes that control cell function, particularly those that control cell growth and reproduction. These cancerous cells develop the potential, over time, to invade or spread to other parts of the body (*metastasis*), a stage at which the patient prognosis becomes terminal.

Breast cancer is one of the more prevalent forms of cancer among the female population globally. Most patients who suffer from breast cancer remain unaware of the disease because there are seldom any physically visible signs of the disease; a fact that holds true for other forms of cancer. For this reason, breast cancer is primarily diagnosed through an annually recommended mammographic screening procedure performed by a radiologist. Mammography is a medical imaging technique that uses low-energy X-Rays (approximately $30KVp$) to capture images of the human breasts. These images, known as mammograms, are then examined by radiologists for the presence of cancerous growths.

Table 1.1: Cancer prognosis five years after diagnosis by prognostic characteristics and age (Five-Year relative survival). (From Mariotto et al. [166])

	All ages (%)		$20-44$ **yrs (%)**		$45-54$ **yrs (%)**		$55-64$ **yrs (%)**		$65-74$ **yrs (%)**		$75+$ **yrs (%)**	
All stages	89	$(89-90)$	88	$(87-88)$	90	$(90-91)$	90	$(90-90)$	91	$(90-91)$	87	$(87-88)$
ER positive	94	$(94-95)$	93	$(92-93)$	94	$(94-95)$	94	$(94-94)$	95	$(94-95)$	96	$(95-97)$
ER negative	79	$(78-79)$	79	$(78-79)$	80	$(79-81)$	80	$(79-81)$	79	$(78-81)$	72	$(70-74)$
Stage I	100	+	98	$(97-98)$	99	$(99-99)$	100	$(99-100)$	100	+	100	+
ER positive	100	+	99	$(98-99)$	100	$(99-100)$	100	+	100	+	100	+
ER negative	97	$(96-97)$	94	$(93-95)$	96	$(95-97)$	97	$(95-97)$	98	$(96-99)$	100	+
Stage II	93	$(92-93)$	92	$(91-93)$	94	$(93-94)$	94	$(93-94)$	93	$(92-93)$	91	$(89-92)$
ER positive	96	$(96-97)$	96	$(95-96)$	97	$(96-97)$	97	$(96-97)$	96	$(95-96)$	96	$(94-97)$
ER negative	84	$(83-85)$	85	$(84-87)$	86	$(84-87)$	86	$(84-87)$	83	$(80-85)$	74	$(70-78)$
Stage Ill	73	$(73-74)$	75	$(74-77)$	78	$(77-79)$	75	$(74-76)$	73	$(71-75)$	58	$(55-61)$
ER positive	81	$(80-82)$	84	$(82-86)$	85	$(84-87)$	83	$(82-85)$	79	$(77-81)$	65	$(62-68)$
ER negative	59	$(57-60)$	59	$(57-62)$	63	$(61-65)$	59	$(56-62)$	58	$(54-62)$	45	$(41-50)$
Stage IV	24	$(23-25)$	32	$(29-35)$	28	$(25-30)$	23	$(21-25)$	23	$(21-26)$	16	$(14-19)$
ER positive	31	$(29-32)$	40	$(36-45)$	35	$(32-38)$	29	$(27-32)$	31	$(28-34)$	22	$(19-26)$
ER negative	16	$(14-17)$	21	$(16-25)$	17	$(14-20)$	16	$(13-19)$	14	$(10-18)$	10	$(7-14)$

The mammographic screening process is not without flaw. Recent studies show the process as being plagued with low sensitivity ($68-92\%$ range), with a notably high type II error rate (false-negative) of 29% in visually detectable cancers [297, 138, 225]. Approximately 50% of these inaccuracies result from human error. While type I errors (false-positives), can have adverse negative impact/effect on the mental health and well-being of the patient, the occurrence of a type II error has a significant impact on the patients prognosis. A missed detection of a cancerous growth in its early stages, during which treatment outcomes result in higher chances of patient survival (as high as 100%), will likely result in a detection during later stages of the cancer (when it begins to manifest physically visible signs), which have a marked lower patient survival rate often as low as 18% (see Table 1.1).

A significant amount of research effort is currently dedicated to addressing these challenges. One body of research focuses on understanding the current diagnostic process and developing tools to improve performance. Other areas of equal importance involve developing new processes that include advanced imaging techniques, improved computer vision algorithms for image understanding, and combining hu-

man beings and computing systems for more efficient and accurate results (computer-aided diagnostic systems).

Understanding the current diagnostic process affords improved patient outcomes in the short run. Intuitively, the mammographic screening process can be modelled as a visual search problem: a radiologist's task is search for a cancerous growth in a mammographic image. This requires an investigation into the radiologist's visual behavior and underlying cognitive process during the screening process. To this end, our research focuses on the application of visual sensory modalities (an eye tracking device), to capture radiologists' visual behavior during the diagnostic screening process. Once captured, we attempt to understand underlying visual and cognitive behavioral processes during the screening process and how these two factors combinatorially affect diagnostic outcomes.

The main area of inquiry for this research is in the development of eye-tracking algorithms to accurately quantify visuo-cognitive behavior of radiologists during the mammographic screening process and ultimately improve diagnostic accuracy in mammography. This research work has four objectives: (i) model radiologists' overall search behavior; (ii) development of spatial and temporal descriptors of visual search behavior during mammographic screening; (ii) evaluation of the efficacy of these features in predicting factors associated with diagnostic performance.

The contributions of the results of this research to the field of computational sciences is the development of eye tracking algorithms for interpreting the behavior or radiologists during mammographic cancer screening and individualized computational models for predicting diagnostic performance. These contributions provide a the foundation for intelligent computing systems that will assist radiologists in managing performance. In addition, intelligent computing systems are applicable in the educational environment to improve training methodology for Radiology residents.

1.1 Overview of Mammography

The early detection of breast cancer affords the one diagnosed with a wider variety of treatment options and an improved chance of survival. There are several medical imaging options available for use in examining human breasts. These include X-Ray imaging, magnetic resonance imaging (MRI), and positron emission tomography scan among others. The most commonly used method for breast imaging is known as mammography. Mammography is a medical imaging technique that uses low-energy X-Rays (approximately $30KVp$) to capture a visual representation of the human breast. These visual representations, which can be film or digital, are subsequently examined by a trained specialist (board certified radiologists) to determine the presence of any cancerous growths.

There are two types of mammographic examinations performed by radiologists: screening mammography and diagnostic mammography. Screening mammograms can be used to check for breast cancer in women who are absent of physically visible signs or symptoms of the disease. A screening mammogram generally involves four views from two X-Ray images of each breast: the craniocaudal (CC) view and the mediolateral oblique (MLO) view [119]. These x-ray images make it possible to detect tumors that cannot be felt. Screening mammograms also capture microcalcifications (tiny deposits of calcium) that sometimes indicate the presence of a cancerous growth.

In contrast, a diagnostic mammogram is administered to a patient who has previously demonstrated abnormality in previous clinical inquiry, such as the presence of a lump, a thickening of the skin of the breast, nipple discharge, breast pain, or other physically visible signs on the breast [119]. These abnormalities may be symptoms of some other disease or benign condition however. Additionally, a diagnostic mammogram may be administered to further evaluate changes found during a screening

mammogram or to view breast tissue when a screening mammogram is otherwise unobtainable due to special circumstances. As with screening mammography, diagnostic mammography also involves four image views from X-Rays images of both breasts. In contrast, diagnostic mammography may offer a more in-depth view at specific areas of the breasts that are of interest.

A large number of research studies have been conducted to assess the impact of mammographic examinations on breast cancer [117]. The adoption of screening mammography for example, has led to an increased early detection rate for breast cancer and a subsequent reduction in morbidity and mortality rates [189]. In a study to determine the efficacy of screening mammography, Kerlikowske et al. performed a meta-analysis on studies reported between January 1966 to October 1993 [134]. They found that the screening mammography significantly reduced breast cancer mortality in women aged 50 - 74 years after 7 to 9 years of follow-up, regardless of screening interval or number of mammographic views used in each screening session.

However, Kerlikowske et al. found no reduction in breast cancer mortality in women aged 40 - 49 years after 7 to 9 years of follow-up [134]. In a more recent study, Narod et al. tracked $50,436$ Canadian women aged $40-49$ years until the age of 60 for breast cancer morality [187]. They found a small but statistically insignificant increase in the cumulative risk of death from breast cancer between women who were assigned annual mammographic screening and women who were not assigned annual screening before age 60. To assess the impact of mammographic screening on breast cancer mortality in Europe, Broeders et al. performed a meta-analytic review of observational studies published on the subject up till February 2011 [39]. They concluded, based on studies where longitudinal individual data were used, that there was a $25-31\%$ breast cancer mortality reduction in women *invited* for screening, and a $38-48\%$ reduction in women who were actually screened [39]. In similar review of

trend studies of breast cancer mortality in Europe, Moss et al. found a range of $28-36\%$ reduction in breast cancer mortality in those studies that compared mortality in time periods before and after the introduction of mammographic screening [184].

Despite the well-studied benefits of screening mammography, the process is not without a few major drawbacks that pose significant challenges to the scientific research community. Most notable among these challenges is the occurrence of error in the mammographic screening process. The *sensitivity* of screening mammography varies because of the multitude of methods used in calculating and reporting [240, 23]. However, a general range for sensitivity in screening mammography generally fall between $68-92\%$ [297, 138, 225].

Human error, both perceptual error and interpretation error, in screening and diagnostic mammography is recognized as a significant problem [147, 144, 19]. Bird et al. performed an analysis of 320 cancers found in a population of women (age range 59 ± 0.3), who had undergone mammographic screening between August 1985 and May 1990 [30]. In this study, they categorized missed lesions (false negatives) as cases where: (a) lesion could be seen in retrospect; (b) undetected by first reader, but subsequently correctly identified by second reader (double reading); and (c) a cancer correctly diagnosed during a mammography examination immediately preceding the pathologic diagnosis, but retrospectively visible, but incorrectly interpreted as negative, on prior mammograms. Bird et al. found that 77 cancers were missed at screening mammography for reasons, which include: having a benign appearance; being present at previous screening; only visible in one of the mammographic views of same breast; being located on the site of a previous biopsy; and, in 47% of the cases, from being overlooked [30]. Similar studies conducted over the past few decades have drawn similar conclusions, which on average show human error accounting for nearly 50% of all diagnostic errors [168, 29, 25, 26, 27].

Over time, there have been notable attempts to improve the sensitivity of diagnostic performance in screening mammography, including the supplementation of mammographic information through complimentary modalities such as ultrasound and magnetic resonance imaging.

Independent double reading of screening mammograms, where two radiologists perform readings of a screening mammogram independently to form a consensus, were proposed and adopted. Many studies have shown the effectiveness of this method in increasing the number of detected cancers [8, 266, 41, 66]. However, this process is associated with an increased workload and cost burden on the radiology community.

To improve image quality, Rangaraj et al. analyzed the effectiveness of an adaptive neighborhood contrast enhancement technique to improve sensitivity in mammographic screening [230]. In this work, Rangaraj et al. evaluated the reciever operating characteristics (ROC) for six experienced radiologists on a corpus of 300 screen-film mammograms consisting of 222 digitized and enhanced mammograms, and 78 unprocessed mammograms. In their results, Rangaraj et al. conclude that radiologists' performance on images with image enhancements was significantly better in comparison with original film and digitized images [230].

Thanks to advancements in, and integration of, imaging and computing technology, computer-aided detection and diagnostic systems (CAD), have been developed to assist radiologists in the image acquisition, presentation, and diagnostic interpretation stages of mammographic screening. Research shows that the use of CAD systems have significantly improved the detection of breast cancer as measured by an increase in the number of cancers detected and increases in radiologists' recall rate 21.2% [74, 38, 183, 265].

1.2 Computer-Aided Detection and Diagnosis of Breast Cancer

Computer-aided detection and diagnosis, in a broad sense, integrates imaging, image processing, computing, machine learning, and artificial intelligence, with the primary aim of improving patient outcomes. The conventional definition for CAD, as a diagnosis that is made by a radiologist who uses the output from an intelligent computing system analysis of a preprocessed medical image as a "second opinion" in the process of detecting lesions and making diagnostic inferences, however, no longer suffices. Over the past several decades, CAD related research interest has grown significantly and, because of the challenging nature of mammography, spurred significant collaborative multidisciplinary research combining the areas of Radiology, Engineering and Computer Science.

While the impact of research developments in CAD are well documented [265, 15, 35, 6, 172], and a number of commercial systems are already available in the United States [231], research and development of new algorithms and technology is still very active. The following subsections highlight the major areas of CAD research.

1.2.1 Computer-Aided Methods for Improved Detection of Masses and Calcification Clusters

Calcifications are an accumulation of minerals (such as calcium) in body tissue, which form as a result of abnormal calcium deposits in soft tissue, causing it to harden. These calcium deposits can be found scattered throughout the tissues of the mammary gland as macrocalcifications and microcalcifications. Microcalcifications (MCs) usually show up on mammograms as small bright spots. The presence of clusters of microcalcifications are an important indicator of malignancy, and they appear in $30-50\%$ of mammographic cases [128, 248]. Since microcalcifications appear brighter than surrounding tissue, a number of methods taking advantage of this

fact have been proposed. One approach to using these properties is the use of image enhancement methods to improve cluster identification. Mass lesions, however, are a specific type of lesion that have volume and occupy a clearly defines space. Masses found during mammographic screening are typically described according to shape, edge characteristics, and density (the number of fat cells present and the density of suspicious cells) [197]. Majority of mass detection algorithms take advantage of the spatial characteristics of a masses as a differentiator from non-masses.

Nishikawa et al. for example, developed a novel processing algorithm to automatically detect microcalcification clusters [191]. Their algorithm involves three main stages: denoising, filtering, and final classification. First, a difference-image filtering technique, using linear filters, is applied to suppress normal anatomical structure of the breast contained within the image, which for microcalcification cluster identification are considered noise. In the second stage, a gray-level threshold based on a full-image histogram is applied, followed by morphological erosion, and finally an adaptive localized gray-level thresholding is used. Resulting in a set of candidate microcalcification cluster locations. A final selection is achieved through a combination of spectral analysis, Minkowski distance based clustering analysis, and image properties (size, shape, pixel intensity etc.). Their algorithm was able to detect 87% of true clusters when tested on a set of 78 mammograms (50% of which contain clusters of microcalcifications) [191].

McLoughlin et al. proposed a noise model based on an estimation of the quantum noise inherent in X-ray imaging [174]. In this work, McLoughlin et al. assume that baseline noise-level in mammograms is as a result of the limited X-ray quanta. Following from this, the quantum noise is estimated as a function of gray level using a square-root model based approach. The local image contrast is improved by decoupling noise–gray level dependencies [174]. More recently , Panda et al. [202] developed

an image processing algorithm to automatically detect microcalcifications and mass lesions. In their work, Panda et al. proposed a three step process of region of interest identification (ROI), two-dimensional wavelet transformation, and feature generation based on Otsu threshold [200], used to automatically perform clustering-based image thresholding [245]. In summary, the methods discussed above apply standard image processing techniques for pre- and post image processing for subsequent detection.

In [304], Yu et al. presented system designed to detect clustered microcalcifications in digitized mammograms. The proposed two-staged system involved the application of statistical and wavelet features for segmentation, followed by detection of microcalcification by a neural network implementation using a total of 31 features extracted from the first step. They report a 90% sensitivity in the detection of microcalcification clusters in a database of 40 mammograms [304].

Campanini et al. developed a novel method in which the mammographic image is encoded as using a multiresolution overcomplete wavelet representation, and subsequently processed through a two-stage machine learning algorithm consisting of a two sequential support vector machines and a final ensemble classifier [43]. In this work, they reported a sensitivity of up to 80% from the DDSM database. Jen et al. [122] developed a two step method for detecting tumorous masses. First, they apply gray level quantization on segmented images for feature extraction, and subsequently they apply principal component analysis to determine weights on each of the features extracted. Jen et al. report a sensitivity of 88% and 86% on two separate datasets. Choi et al. [47] proposed a novel computer-aided detection framework to improve sensitivity of mass detection by combining an unsupervised and supervised machine learning algorithms to improve identification of regiouns-of-interest (ROI), combined with an ensemble classifier. Their results suggest a 70% reduction in false-positive detections, but a 4.6% loss in sensitivity [47].

1.2.2 Computer-Aided Detection of Architectural Distortions and Bilateral Asymmetry Anomalies in Mammograms

A second general area of research in computer-aided detection and diagnosis of breast cancer is in the identification of structural abnormalities in anatomy of the breasts, which are not physically conspicuous as is the case with microcalcifications and mass lesions. The first type of anomaly in this group are architectural distortions. Architectural distortions rank as the third most common sign of non-palpable breast cancer found in mammograms [229], which due to subtlety and apparent arbitrary physical characteristics, are often missed during mammographic screening [136]. According to the Breast Imaging Reporting and Data System (BI-RADS), architectural distortion is defined as a distortion in the normal architecture of breast in the absence of a physically visible mass [198]. "This includes spiculations radiating from a point and focal retraction or distortion at the edge of the parenchyma" [198].

Architectural distortion accounts for $12-45\%$ of missed breast cancers in screening mammography [297]. In a review of 234 screen-detected and interval cancer cases (aged $44-84$ years) of screening and diagnostic mammogram cases between 1991 and 1996, Broeders et al. concluded that detection of architectural distortion and non-spiculated high-density masses can lead to an improvement in the prognosis of breast cancer patients [40]. Deducing from this, there have been a number of attempts to characterize and quantify architectural distortion. Guo et al. investigated the use of Hausdorff fractal dimensions with an support vector machine classifier to characterize architectural distortions [81]. They achieved an accuracy of 72.5% in identifying architectural distortions in a set of 40 regions of interest (ROIs). Tourassi et al. used fractal dimension to characterize architectural distortions patterns in ROIs. The area under the receiver operating characteristics (ROC) curve was 0.89 on a dataset of

1500 ROIs (112 with architectural distortions).

The second type of anomaly is generally referred to as bilateral asymmetry. Radiologists use a perceived difference in symmetry between the left and right mammograms of a given patient as an indicator to diagnose breast cancer [69]. According to BI-RADS bilateral asymmetry indicates a difference in volume or density of breast tissue in the absence of a distinct mass, or more prominent ducts, in corresponding areas between the left and right breasts of the same patient [198]. In a study of 252 asymptomatic women who had normal mammography but went on to develop breast cancer [243], Scutt et al. found measures of bilateral asymmetry to be strong predictors of breast cancer. In [181], Miller and Astley new methods to detect bilateral asymmetry by comparing pairs of corresponding anatomical structures, which are detected using automatic segmentation of breast tissue types. In a more recent work [44], Casti et al. investigated differences structural information of automatically detected regions using spherical semivariogram descriptors and correlation-based similarity metrics in spacial and wavelet domains. Using features extracted from gray-scale values and magnitude and phase response of Gabor filters, they evaluate the performance of their method with linear discriminant analysis, Bayesian classifier, and artificial neural network with radial basis functions, on 188 two-view mammograms. They reported an accuracy of up to 0.94, and a sensitivity and specificity score of 1 and 0.88 respectively.

1.2.3 Computer-Aided Detection For Real-Time Support in Mammography

In this section, we discuss alternative applications of computer-aided detection research in mammography. Computer-aided detection methods discussed to this point remain decoupled from the mammographic screening and diagnostic process, where computing methods are applied before or after the mammographic process to verify

or question diagnostic decision. In contrast, some research studies have proposed the use of computer-aided systems in parallel to enhance the mammographic process. Karssemeijer et al. [129] developed an interactive mammographic screening system, which enabled the image reader to probe regions of interest on the image for relevant CAD information. Using a corpus of 60 cases, they tested their system on two radiologists and four non-radiologists and found a significant increase in performance, from 28.27% without CAD to 38.03% with the use of the interactive CAD system. Samulski et al. developed a similar customized workstation in which readers were able to probe mammographic image locations for relevant CAD information [238]. They tested their system with four screening radiologists, who were asked to review 120 cases. A significant ($p = 0.012$) improvement in detection performance was reported with an improvement in average sensitivity from 25.1% without CAD, to 34.8% in CAD-assisted sessions.

Content-based image retrieval methods have been applied as diagnostic tools to aid radiologists by providing data from previous cases, which are relevant to a current case based on computed similarities of image content. Qi and Snyder applied simple image pixel properties (including shape, size, and intensity) to quantify similarity between images [264]. Tourassi et al. [269] evaluated the use of information-theoretic image similarity measures in content-based retrieval and detection of masses in screening mammograms. They compared the precision and detection accuracy of eight entropy-based similarity measures on a database of 1820 mammographic ROIs. Their results showed image similarity measures can be used for semantic similarity, and the presence of a masses.

1.3 Image Perception Research in Screening Mammography

In previous sections, we introduced research work in computer-aided detection in mammography. To summarize, the class of research discussed focused on computational understanding of mammograms for the purpose of improving performance outcomes in mammography. A second important class of research focuses instead on computational understanding of the observer component for the purpose of improving diagnostic performance in mammography [142]. We define diagnostic performance of the human in mammography as a measure of how well the subject is able to predict the presence or absence of a cancer, and for the latter, how well the extent or magnitude of the disease or condition is measured. Deducing from this, we can conclude that perception and cognition serve as the primary drivers of the diagnostic process, and ultimately predict diagnostic performance. Therefore to understand and improve diagnostic performance (i.e. to reduce human error), it follows that we must examine both the perceptual and the cognitive processes during mammography. Since the two sources of human error screening mammography: perceptual error and interpretation error, account for nearly 50% of missed diagnosis [147, 19], a large body of research has been directed toward understanding how these errors occur and applying this knowledge to provide feedback during the diagnostic process, and to improve training methodology.

Nodine et al. [193] investigated the correlation between number of years of experience, training, and mammography expertise. In this study, they investigated the effect of perceptual and cognitive skills in both detection and interpretation in mammography by analyzing the performance of three categories of image readers (expert radiologists, Radiology residents, and mammography technologists). Nodine et al. found that experts had the best performance, and residents had a significantly

lower performance, which however, was equivalent to that of mammography technologists. They concluded that poor performance by residents resulted from a lack of perceptual-learning experience during mammography training, and recommended a systematic *mentor-guided* training and feedback to improve image perception and decision making [193]. The use of specialized technologies, such as eye tracking devices, to record eye position position data during mammography, enable us to develop psycho-physiological models useful in understanding how visual and cognitive errors occur, and ways to reduce their occurrence.

1.4 Eye Tracking

Eye tracking refers to a process of measuring the movement of the eye relative to the head. Eye tracking devices are bio-sensors that measure psycho-physiological response through changes in configuration of the human eye. These changes include positional measures such as visual fixation, saccadic movements, and scanpath, and non-positional measures such as blinks and pupil dilation and constriction. Using data from eye-tracking sensors, we can model human perception, cognitive processes, and responses to external stimuli.

Eye tracking technology has experienced improvement in precision, capability, and affordability. This trend has fostered the popularity of eye tracking both as a research tool and an interaction modality in a large variety of disciplines. Furthermore, eye tracking is one of the best noninvasive methods which provide a window into users' visual and cognitive processes relating to response and intent [64] [163].

1.4.1 Visual Perception

The human eye perceives light rays through the cornea. the cornea is the clear, transparent front covering which, acting as a lens, admits light and begins the refractive process. It also keeps foreign particles from entering the eye. The pupil is an

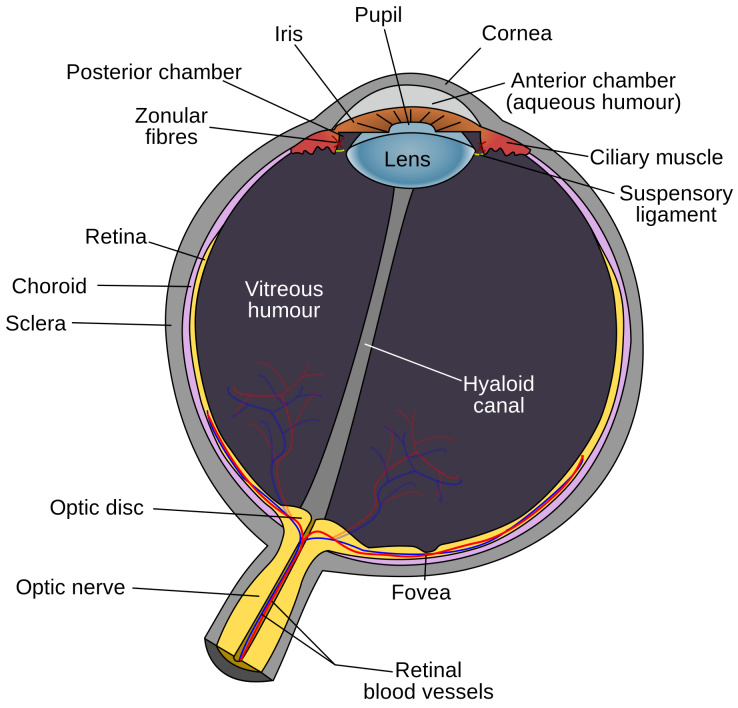


Figure 1.1: An illustration of the human eye (From Rhcastilhos [53]).

adjustable aperture in the center of the iris that controls the intensity of light permitted to strike the lens. The iris turns the image upside down in the lens and then projects it onto the back of the eyeballthe retina. The lens focuses light through the vitreous humor, a clear gel-like substance that fills the back of the eye and supports the retina.

The retina is filled with light-sensitive cells, called cones and rods, which transduce the incoming light into electrical signals sent through the optic nerve to the visual cortex for further processing. Cones are sensitive to high spatial frequency (visual detail) and provide color vision. Rods however, are very sensitive to light, and therefore support vision under dim lighting conditions

A small area at the bottom of Figure 1.1, known as the fovea, which spans less than 2° of visual field, has an extremely dense concentration of cones, while they are very sparsely distributed in the periphery of the retina. This results in our having full acuity only in this small area, roughly the size of a thumb nail at arms length. This has the implication that in order to see a selected object sharply, like a word in a text, we therefore have to move our eyes, so that the light from the word falls directly on the fovea; a process known as foveating. Only when we foveate on a word or object can we read or see it sharply.

For video-based measurement of eye movements, both the pupil and the cornea are very important. Though less known, the cornea covers the outside of the eye, and reflects light. The reflection visible in a person's eyes comes from the cornea. When tracking the eyes, we are interested in a single reflection. Because visible light produces numerous reflections, to avoid all natural light reflections, we record in infrared by illuminate the eye with one (or more) infrared light sources.

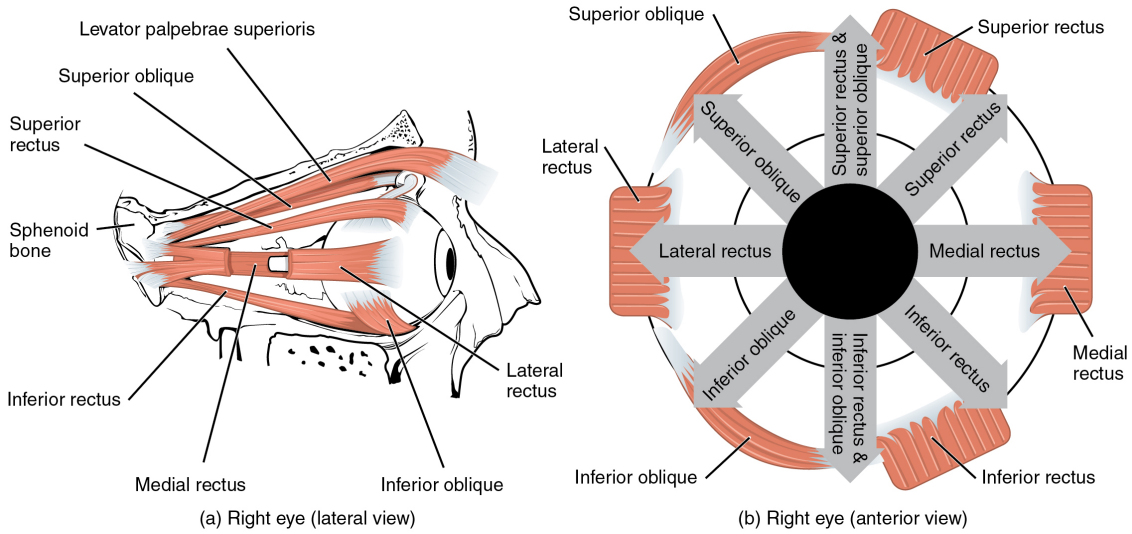
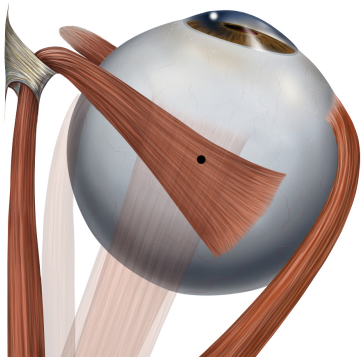


Figure 1.2: An illustration of the human eye muscles that generate the vertical up-down movements (superior and inferior rectus), the horizontal outward-inward movements (lateral and medial rectus), and the torsional rotating movement (superior and inferior oblique)(From OpenStax College, [52]).

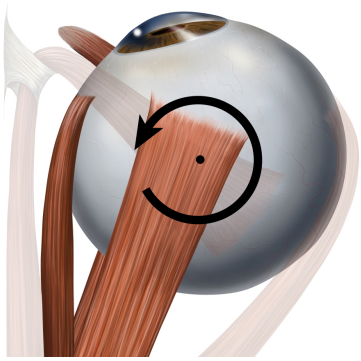
1.4.2 Eye Events

Eye events refer to voluntary or involuntary change in the configuration of the eye, which may or may not constitute actual movement of the eye (e.g. pupil dilation or constriction), but which help the subject to acquire, fixate or track visual stimuli. The eyes are able to move through the coordinated activity of a system of six muscles (see Figure 1.2.

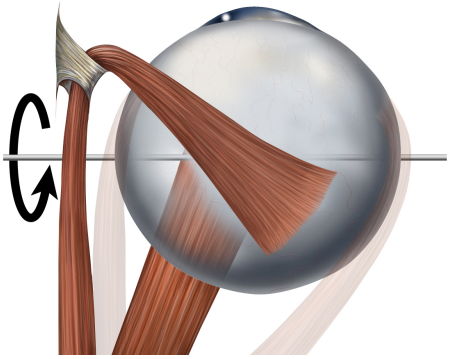
The movement of the human eye is controlled by three pairs of muscles, depicted in Figure 1.2. The combined and coordinated actions of these muscles (depicted in Figure 1.3) are responsible for horizontal (yaw), vertical (pitch), and torsional (roll) eye movements, respectively, and hence control the three-dimensional orientation of the eye inside the head. According to Donders law [273], orientation uniquely decides the direction of gaze, independent of how the eye was previously orientated. Large



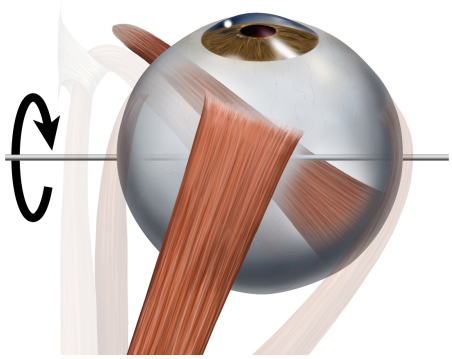
(a) Eye movement of lateral rectus muscle (From Lynch [157]).



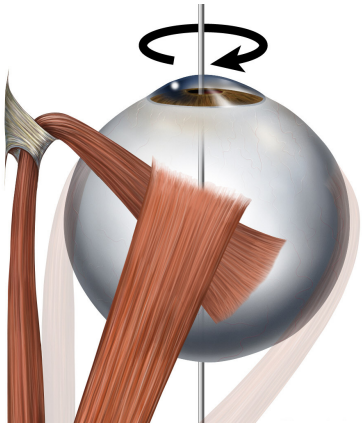
(b) Eye movement of medial rectus muscle (From Lynch [158]).



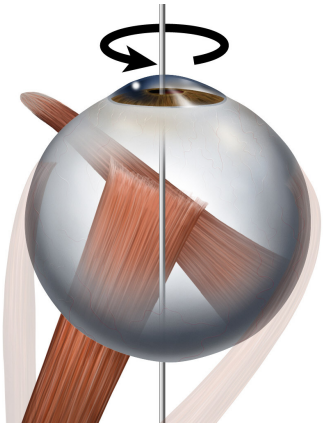
(c) Eye movement of inferior rectus muscle (From Lynch [159]).



(d) Eye movement of superior rectus muscle (From Lynch [160]).



(e) Eye movement of superior oblique muscle (From Lynch [161]).



(f) Eye movement of inferior oblique muscle (From Lynch [162]).

Figure 1.3: Superior view of muscles responsible for horizontal (yaw), vertical (pitch), and torsional (roll) eye movements.

sections of the brain control these muscles to direct the gaze to the desired locations in space.

Humans and other primates (including other vertebrates) primarily engage in seven types of voluntary and involuntary eye movement: fixation, saccade, glissade, smooth pursuit, microsaccade, tremor, and drift (see Table 1.2) [115]. A *fixation* refers to a state where the eyes remain still (within a small radius) over a period of time, such as is the case when the eyes pause on a given word while reading. Fixating on a point or region is generally considered as a measure of attention to a given position or region of interest, even though this is not always the case. While there is no universally excepted method for detecting fixations, there are established parameters based on ocular physiology, which permit a reasonable criteria for detecting and extracting fixations from gaze data. A typical algorithm to determine a fixation event uses the mean X and Y eye position coordinates measured over a minimum period of time during which the eye does not move more than some maximum amount. This algorithm requires that a point-of-gaze must continuously remain within a small area (approximately within 1-degree visual angle) for some minimum amount of time (approximately 100ms).

The eye is not completely still during a fixation, but exibits three distinct types of micro-movements: *tremor*, *microsaccades*, and *drifts* [169]. A tremor is a small movement of approximately 90 Hz. The exact role of tremors is still a subject of research; it is generally believed to be imprecise muscle control. Drifts are slow movements that shift the eye away from the centre of fixation, while the counter movement, a microsaccades, serves to quickly return the eye back to the center of fixation.

The rapid motion of the eye from one fixation to another, from word to word while reading, for instance, is called a *saccade*. Saccades are considered the fastest movement the body can produce; typically taking 3080 ms to complete. It is a gen-

Table 1.2: Basic measures of positional eye movement events.

Description	**Duration (ms)**	**Amplitude**	**Velocity**
Fixation	$200 - 300$	N/A	N/A
Saccade	$30 - 80$	$4 - 20°$	$30 - 500°/s$
Glissade	$10 - 40$	$0.5 - 2$	$20 - 140°/s$
Smooth Pursuit	N/A	N/A	$10 - 30°/s$
Microsaccade	$10 - 30$	$10 - 40'$	$15 - 50°/s$
Tremor	N/A	$< 1'$	$20'/s$ (peak)
Drift	$200 - 1000$	$1 - 60'$	$6 - 25'/s$

erally held view that human beings are perceptively blind during most of a saccadic event; a phenomena illustrated in Figure 1.4.

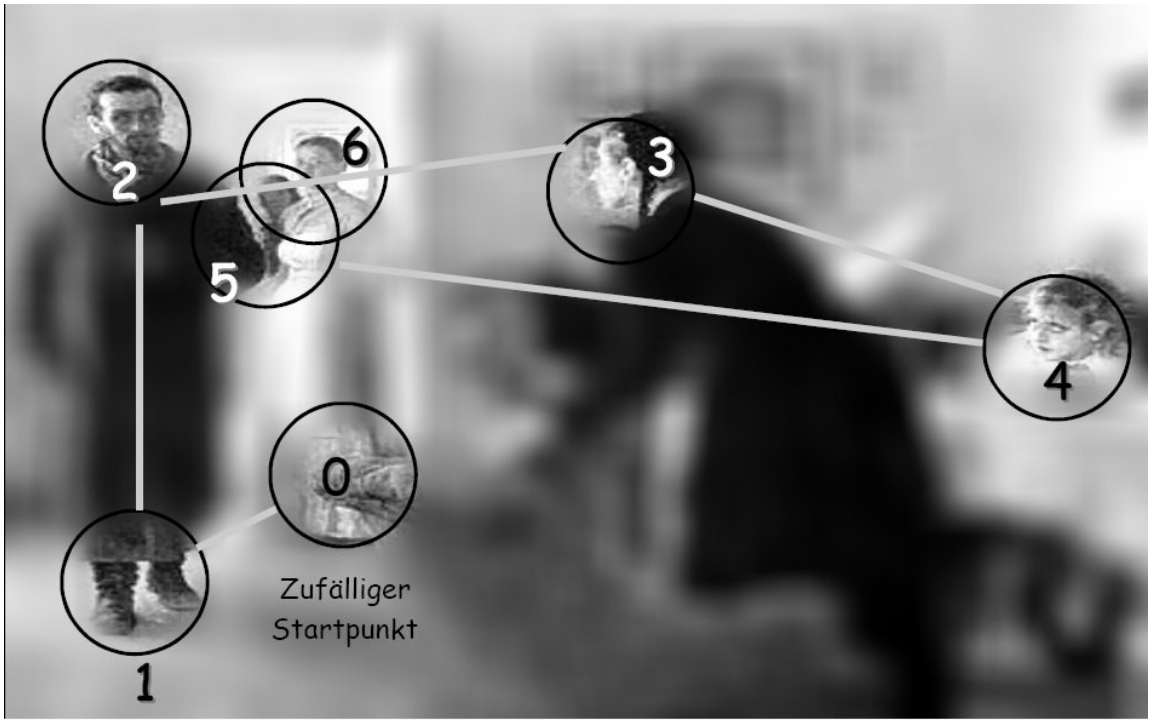


Figure 1.4: Eye movements during the first 2 seconds of viewing a picture. (based on data from Yarbus [298]).

An important characteristic of saccades is that they rarely take the shortest path between two points, but instead undergo one of several *shapes* and *curvatures*. Since a saccade is described in terms of the gaze data between detected fixations, a saccadic event can be computed as gaze points connecting the completion of one fixation to the beginning of the next fixation. The saccadic movement is not mechanically precise, that is they do not stop directly at the intended target, but instead the eye wobbles before coming to a stop. This post-saccadic movement is referred to as glissadic movements or *glissade*.

The movement of the eye is characteristically different in the case of following or tracking a moving object such as a bird flying across the sky. This type of eye movement is usually slower and referred to as *smooth pursuit*. The difference between the saccadic and the smooth pursuit movements is that the latter is driven by and requires a moving target, while the former can be made independent of any visiual stimulus. The term *scanpath* refers to the path (in two-dimensions) of the eye as it moves through a spatial scene. It can be described as the route of oculomotor events through space within a certain timespan. This definition assumes that the path has a beginning and end, and therefore a length. The scanpath can be characterized in one of two ways: (1) the raw gaze scanpath, and (2) the fixation-based scanpath. The raw gaze scanpath refers to all physical movements of the eye within a certain timespan, which encompasses saccades, glissades, microsaccades, etc. However, the fixation-based scanpath is the path formed when all fixations within a certain timespan are connected through a straight line.

1.4.3 Eye Tracking Research in Radiology

There is a wealth of research in the domain of radiology, specifically mammography, in which eye tracking sensors are used as a tool to gain a better understanding of

visual behavior during the mammographic process. Most research on visual perception in mammography fall under one of two categories: (1) expertise, or (2) diagnostic error.

Research on visual perception and expertise in mammography focuses on how expertise is developed and attempts to characterize visual behavior associated with expertise. This area of research uses eye tracking to measure and quantify observable factors, which differentiate Radiology experts, such as duration and scanpath, from non-experts.

In a seminal work analyzed visual behavior during the reading of chest radiographs [150], Kundel and La Follette examined eye movements of a total of 24 subjects (including untrained laymen, medical students, Radiology residents, an staff radiologists) for trends that correlated with experience and expertise. Kundel and La Follette found an evolution of search patterns from a localized central pattern of untrained laymen, to a more circumferential pattern exibited by experienced staff radiologists, based on visual observation of the scanpath. Although Kundel and La Follette reported visibly noticeable differences in the scanpath with an increase in experience and expertise, the aggregate measures of fixations fail to characterize these differences.

In a study, analyzing scan strategies in mammography [139], Krupinski examined the eye-event data of six image readers (3 staff mammographers and 3 Radiology residents) for differences in the search behavior of experienced and inexperienced radiologists during a mammographic reading. Using measurements of gaze duration and dwell time, Krupinski found differences between experienced and inexperienced image readers, reporting that more experienced image readers had shorter dwell times and longer gaze duration. However, the findings reported in [139] utilize simple aggregate eye-event features on a relatively small population and corpus of mammographic

images. Krupinski also reported that although there were notable differences between groups, these differences were not statistically significant.

To investigate human factors associated with the proficiency of diagnostic pathology, Krupinski et al. [143] conducted a study examining the eye movements of nine image readers, belonging to one of three experience groups (medical students, Pathology residents, and pathologists). In this study, each of the nine slide readers took part in a single 45 minute virtual slide reading session, during which they examined 20 breast core biopsy slides while their eye movements were recorded. They found that experienced pathologists had the longest saccade length on average (measured in seconds) compared with residents, who in turn had longer saccade lengths on average compared with medical students.

In addition, they found that the average saccade velocity (measured as length per second) for experienced pathologists was significantly lower than residents, who's average saccade velocity was higher than the average velocity for medical students. They also reported that the decreasing trend in saccade velocity with years of experience was consistent within the experienced pathology group, with the most experienced pathologist having a significantly lower average saccade velocity than the less experienced pathologists. The findings by Krupinski et al. suggest distance and velocity measures of eye movements during visual search may also be important factors in differentiating between experienced and inexperienced image readers. These findings are also based on simple aggregate eye-event features, which require *a priori* knowledge about the image stimulus (regions of interest).

Manning et al., examined groups of image observers of varied levels of expertise to investigated the influence of training and experience on visual search behavior and diagnostic performance during the interpretation chest nodule [164]. Eye-event data of four observer groups: eight experienced radiologists, five experienced radiog-

raphers before and after six months training in chest image interpretation, and eight undergraduate radiography students were recorded during detection and localisation of significant pulmonary nodules in posterior-anterior views of the chest. Manning et al. analyzed measures of fixation, saccade, coverage, and gaze duration, and reported finding experienced radiologists have significantly longer saccadic amplitude, a fewer number of fixations, and shorter duration of gaze than all other groups. They also reported the experienced group of radiologists and radiographers after training having a better detection performance, as measured with an Alternate Free Response Operating Characteristic technique, than the remainder groups. While Manning et al. observed distinct differences in search strategies between the experienced and inexperienced observers, there was no direct link between eye-event features used and task performance.

In a more recent work, Krupinski et al. examined and characterized visual behaviour of four pathology residents as they progressed through residency training (at the beginning of their first, second, third, and fourth years of residency) [141]. Krupinski et al. recorded eye-event data for each resident while viewing a series of 20 digitized breast biopsy whole slide images at each of the above mentioned stages in their residency training. During the experiments, each resident was tasked to examine the image and select three areas that they would want to zoom in on for further examination and diagnostic detail at a higher resolution. Using numberosity and movement measures of fixation, saccade, dwell time, and region of interest (ROI), Krupinski et al. reported finding that as the residents progressed through residency training, there was a decrease in the overall time taken to make decisions about where to zoom, a significant decrease in total fixations, and a decrease in time spent examining areas with no diagnostic value; indicating improved overall search efficiency. The findings presented in [141] are significant in establishing that visual

behavior changes as residents progress through training and gain more experience. However, the study was based on a small population ($x = 4$). The important findings in this study were also based on simple eye-event features, which require *a priori* knowledge about the image stimulus (regions of interest).

In a study on breast cancer detection, Tourassi et al. investigated the relationship between radiologists gaze, diagnostic decision, and image content of mammograms during mammographic cancer screening [267]. They examined eye-event data from six image readers (three breast imaging radiologists and three Radiology residents), and image content of 20 screening mammograms. Using machine intelligence algorithms, they developed predictive models combining image content, visual behavior, cognition, and risk of diagnostic error. Their results suggest that machine learning can be utilized in combination with image content and the image reader's visual behavior to develop user-dependent models for predicting risk of diagnostic error in breast cancer lesion detection and characterization with an accuracy of 59%. The findings reported by Tourassi et al. [267] indicate a major step in linking image content, human perception, human cognition, and human error in mammographic breast cancer detection. However, this work is based on a small population ($x = 6$) and a very small sample size ($n = 20$). In addition, the predictive models generated in this work are also dependent on *a priori* knowledge about the image stimulus (regions of interest).

Although many investigators have examined visual search behavior in the radiology, and more specifically mammography, the features utilized in these investigations have hitherto been limited to: (1) simple aggregates of positional, duration, or numerosity measures of eye-events; (2) based on *a posteriori* knowledge of regions of interest within the stimulus image.

While these features provide informative insights into visual search behavior dur-

ing the mammographic screening process, they do so in a limited fashion. They also fail to capture observable differences such as the differences in gaze path trajectory between experienced and inexperienced image readers as for example is reported in [164], and are therefore insufficient to fully characterize the visual search process.

In this work, we developed novel eye-event primitives and extended existing algorithms, inspired from other domains including sketch recognition, data mining, and information retrieval to improve prediction of mammographic case characteristics (such as case pathology and breast parenchyma density), radiologists characteristics (including individual identity and experience level), and risk of diagnostic error.

We have developed an eye-event feature, fractal dimension, which requires no prior knowledge of regions of interest in stimulus image. The fractal dimension, which characterizes the space filling capacity of a pattern, provides a statistical index of complexity of radiologist's scanpath during mammographic screening. This index, we hypothesize, can accurately characterize the characteristics of a mammographic case, the image reader's identity and experience, and the risk of diagnostic error. Bhat and Hammond have shown a related measure, Shannon's entropy, to be a good measure to distinguish between text and shape in sketch recognition [28].

We implemented an existing machine learning algorithm, time series shapelet analysis, originally developed for the data mining and information retrieval research domain, and applied it on eye-event data recorded during mammographic screening. In addition, we extended the time series shapelet algorithm in a manner that optimizes its utility for eye tracking data.

We applied sketch gesture recognition techniques to extract geometric-based features from eye-event data to characterize visual search behavior during the mammographic screening process. These features, we hypothesize, provide a more fine-grained characterization of scanpath by aggregating the spatial (shape), directional,

and kinetic properties of its constituent saccadic movements. We compared features described above with a previously developed method, which applies timeseries shapelet analysis to extract discriminative information to from changes in pupil dilation from eye-event data.

2. FRACTAL ANALYSIS OF RADIOLOGISTS VISUAL SEARCH BEHAVIOR IN SCREENING MAMMOGRAPHY*

The goal of this study was to test the efficacy of radiologists visual search complexity, computed using fractal dimension, as a predictor of image characteristics, case pathology, and image reader experience level when viewing 4-view mammographic cases, as they typically do in clinical practice. The study was performed for the task of mammographic screening as typically done in clinical practice. Eye-tracking data and diagnostic decisions for 100 mammographic cases, collected from seven Radiology residents and three board certified radiologists, formed the corpus used for this study. Visual search complexity, using gaze data extracted from eye-tracking data, was estimated using fractal dimension computed using the Minkowski-Bouligand box-counting method. Mass conspicuity, assessed according to the subtlety rating, and parenchymal density, were provided in the DDSM truth files. Individual factor and group-based interaction ANOVA analyses were investigated.

The characteristics of a mammographic case, including case pathology and breast parenchymal density, image reader experience level, and individual differences each factor as independent predictors of a radiologists visual scanning pattern complexity in screening mammography. No higher order effects were found to be significant.

An aggregate characterization of visual search behavior, captured in visual search complexity, is dependent on case properties and image reader characteristics.

*Description of methods and experimental results are reprinted with permission from "Fractal analysis of radiologists' visual scanning pattern in screening mammography," by Folami T Alamudun, Hong-Jun Yoon, Kathy Hudson, Garnetta Morin-Ducote, and Georgia Tourassi, 2015. *Proceedings of SPIE*, 9416, pp. 94160T-94160T-8, Copyright 2015 by SPIE.

2.1 Introduction

Breast cancer is the most frequently diagnosed form of cancer and the second leading cause of cancer-related deaths among women worldwide. The mortality rate for this disease is largely dependent on early diagnosis through mammographic screening [80]. Statistics show that through early detection through mammographic screening, while the disease is localized, patients have a 98.5% relative survival rate versus 25% when the cancer is metastasized, a point at which the disease becomes incurable [247].

However, previous studies show the mammographic screening process is susceptible to different types of error resulting in misdiagnosis, with 50% of misdiagnosis resulting from human visual error [30, 297, 25, 26]. The topic of diagnostic error has received a lot of attention in recent years. To this end, the medical research community has focused on the perceptual and cognitive processes related to decision making to better understand the causes of error. In radiology, misdiagnosis is attributed to visual search and interpretation errors [31, 140].

For over half a century, a large number of studies have focused on the radiologists visual scan pattern during the image reading process. Findings from these studies indicate prevalence of errors in two general categories: (1) how radiologists find what they are looking for (visual search); and (2) how radiologists interpret what they are looking at (image interpretation) [139, 175, 176, 177, 178, 186]. A large body of eye-tracking research has also focused on gaining a better understanding of the relationship between visual search and diagnostic decision by analyzing radiologists eye movements recorded during the diagnostic process 15–21 [62, 145, 146, 150, 147, 149, 165].

In a study of scanning strategies in mammography [139], Krupinski found differ-

ences between experienced and inexperienced image readers when comparing dwell times extracted from eye position data gathered while viewing mammographic cases. Kundel et al. investigated the occurrence of a global perceptual process, as evidenced in the saccadic movements during the initial viewing of an image, in the analysis of a mammographic image and its importance in the identification of abnormalities [148]. They found that more experienced radiologists develop this global perceptual process as a search strategy than their less experienced counterparts [148].

A More recent research study by Voisin et al. showed the efficacy of eye-tracking in predicting diagnostic performance [281]. Voisin et al. conducted laboratory studies and applied machine learning techniques to predict error during the diagnostic characterization of mammographic lesions by combining features from radiologists gaze behavior, and textural image characteristics [281]. In a related study on breast cancer detection, Tourassi et al. investigated the relationship between radiologists gaze, diagnostic decision, and image content of mammograms during mammographic cancer screening [267]. Their results suggest that machine learning can be utilized in combination with image content and the image reader's gaze characteristics to develop user-dependent models for predicting errors in breast cancer lesion detection and characterization.

Although many investigators have examined radiologists visual scanning patterns for screening mammograms, the discovered patterns are typically summarized with respect to features such as total time examining a case, time to initially hit true lesions, total dwell time, number of hits, etc. While informative, these features fail to capture the gaze path trajectory and therefore they cannot fully capture the complexity of the visual search process. In addition, earlier studies were based on single view mammograms, which is not consistent with clinical practice. Mammographic screening entails simultaneous viewing of 4 coordinated breast views. The purpose

of this study was to address the limitations of the earlier investigations and attempt to characterize the complexity of the radiologists visual search activity when viewing 4-view mammographic cases as a function of three factors: (i) breast parenchymal density, (ii) case pathology, and (iii) radiologists experience level. Our study focuses primarily on mass detection, which is associated with a higher detection error rate than microcalcifications [30, 297, 26, 186].

2.2 Materials and Methods

2.2.1 Image Database

Table 2.1: Specifications of the 100 four-view screening mammograms used in the study.

Ground Truth	**Patient Age**	**Breast Density**	**Mass Subtlety**	**Total Abnormalities**	**No. of Cases**
Normal	Range: 3668 (56.2 $\pm$ 10.6)	Range: 14 (Median: 2)	N/A	N/A	25
Benign	Range: 3482 (56.9 = 13.4)	Range: 13 (Median: 2)	Range: 35 (Median: 5)	Range: 13 (Median: 1)	25
Malignant	Range: 3783 (64.3 $\pm$ 12.4)	Range: 14 (Median: 2)	Range: 15 (Median: 5)	Range: 13 (Median: 1)	50

To perform this study, 100 screen-film mammograms were selected from a corpus of mammographic cases digitized with a high resolution LUMISYS scanner (50m per pixel, 12 bit) from the University of South Floridas Digital Database for Screening Mammography (DDSM) [110]. Each DDSM case contains 4 images, the craniocaudal (CC) and mediolateral oblique (MLO) view images of both the left and the right breasts as well as associated ground truth established via biopsy, additional imaging, or 2-year follow-up, radiologists assessment using the BI-RADS^{TM} lexicon [198], and patient age.

Each of the 100 cases was manually selected to cover a broad range of mass margin and shape characteristics. Of the 100 cases selected, 50 included biopsy-proven malignant masses, 25 cases included biopsy-proven benign masses, and the remaining 25 cases were normal as determined during a 2-year cancer-free follow-up patient evaluation. Therefore, all mass cases selected for the study included clinically actionable masses. Mammograms with masses deemed as benign-without-callback were excluded. The overwhelming majority of the mass cases (72 out of 75) did not include any microcalcifications. Mass conspicuity was assessed according to the subtlety rating provided in the DDSM truth files. These ratings ranged from 1 (suggesting a subtle lesion) to 5 (suggesting an obvious lesion). A complete list of the DDSM cases used in this study is provided in the Appendix. Table 2.1 provides details on the selected cases, including information on the patients age and breast parenchymal density. The parenchymal density is also provided in the DDSM truth files, and it ranged between 1 (fatty) to 4 (dense), according to the BI-RADS^{TM} lexicon [198].

2.2.2 Data Collection Protocol

Ten readers of variable experience levels from an academic institution were recruited to conduct blind review of the selected mammograms (see Table 2.2). Each reader was asked to report the location of any suspicious mass and provide a corresponding BI-RADS rating as typically done in clinical practice. Of the ten readers, three were experienced MQSA-certified radiologists each with at least nine years of dedicated mammographic experience, four radiology residents with at least three mammography rotations, and three radiology residents with at most two mammography rotations (see Table 2.2). Institutional review board approval was obtained prior to the study. Human subject recruitment and data collection was done according to a protocol approved by the Oak Ridge Site-Wide Internal Review Board. All

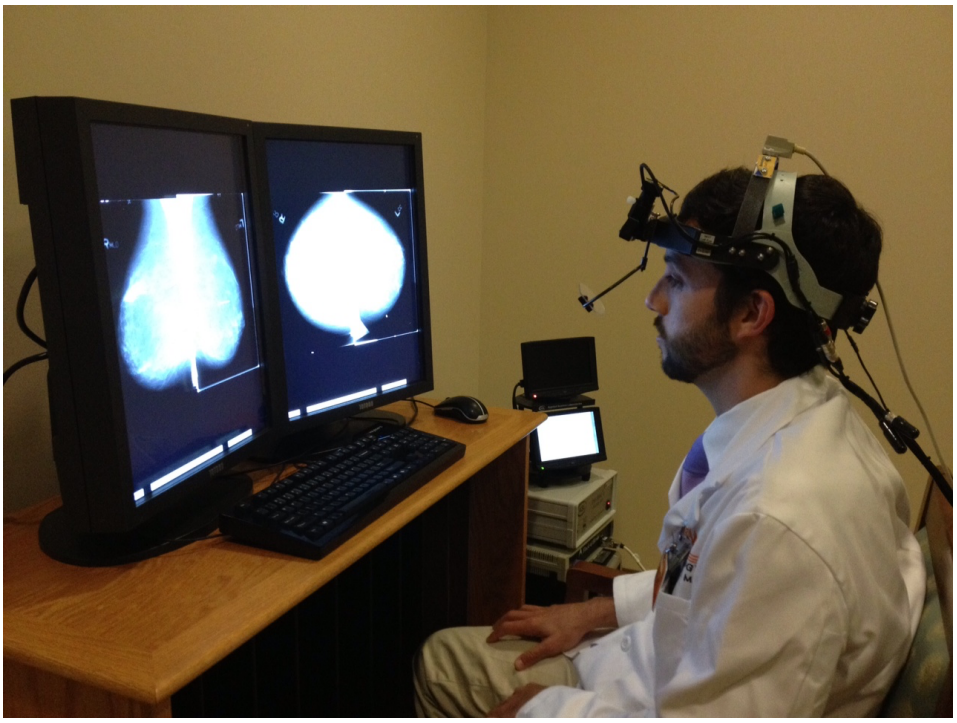
participants signed an informed consent form.

Table 2.2: Summary of characteristics of study participants.

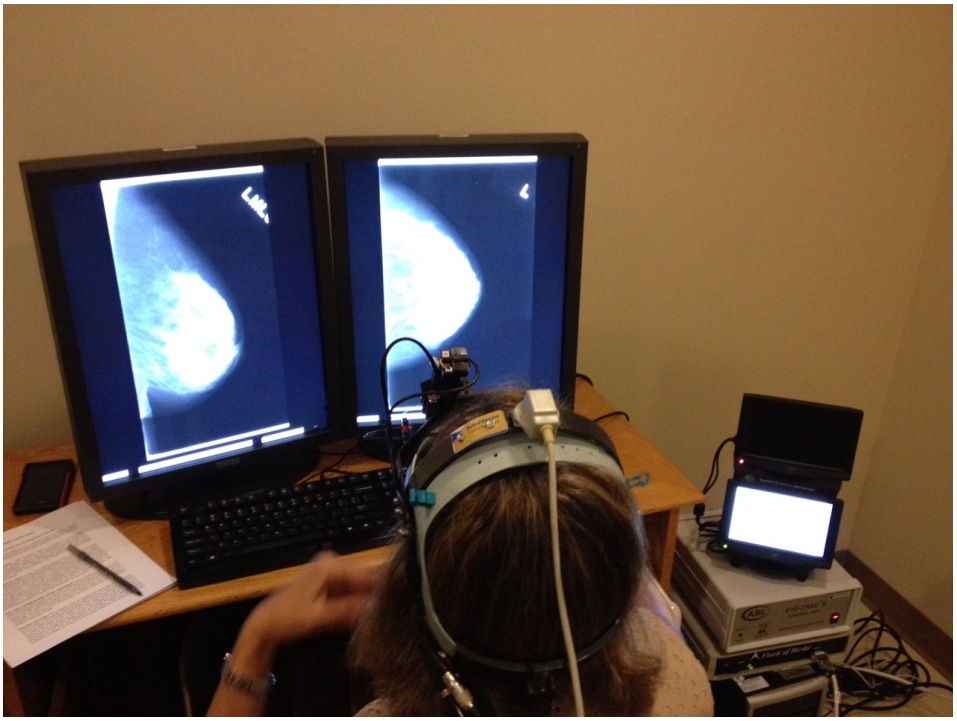
Reader Type	**Experience Level**	**No. of Participants**
Expert Radiologist	$>$ 10 yrs of practice	2
Expert Radiologist	$<$ 10 yrs of practice	1
Advanced Resident	$>$ 2 mammo rotations	4
New Resident	$\leq$ 2 mammo rotations	3
Total		10

A customized graphical user interface (GUI) was developed in-house for study participants to view each mammographic case and record their findings. Two medical grade monitors were used (dual-head 5MP mammo-grade Totoku LCD monitors calibrated to the DICOM display standard). The four mammographic views (LCC, RCC, LMLO, RMLO) were initially displayed at low resolution (two views per monitor) to fit the screen. To assess breast symmetry, the users could select the MLO views to be displayed on the left monitor and the CC views to be displayed on the right monitor (Figure 2.1a). The readers were also able to select and view a single breast at full spatial resolution with the MLO view displayed on the left monitor and the CC view displayed on the right monitor (e.g., Figure 2.1b). Table 2.3 enumerates all possible "hanging protocols" implemented in the GUI.

Please note that based on the allowable protocols RMLO could never appear on the right monitor while LCC could never appear on the left monitor. In addition, the GUI provided the functionality of zooming in/out, panning, and magnifying glass for detailed reading of each mammographic view. During the reading sessions, each reader was outfitted with an H6 head-mounted eye-tracker, with a 60 Hz sampling rate, and eye-head integration from Applied Science Laboratories (ASL, Bedford,



(a) Dual Display showing default arrangement.



(b) Dual display showing MLO and CC views of the right breast.

Figure 2.1: Image reader outfitted with eye-tracking apparatus reviewing a mammographic case.

Massachusetts, USA). The eye-tracker recorded each readers eye position data to within 0.5° of accuracy.

Table 2.3: Possible configurations for a combined two-dimensional data representation.

	Left Monitor		**Right Monitor**	
Label	**Left Image**	**Right Image**	**Left Image**	**Right Image**
C1	Right (RMLO)	Right (RCC)	Left (LMLO)	Left (LCC)
C2	Right (RMLO)	Right (RCC)	Left (LCC)	Left (LMLO)
C3	Right (RCC)	Right (RMLO)	Left (LMLO)	Left (LCC)
C4	Left (LMLO)	Right (RMLO)	Right (RCC)	Left (LCC)
C5	Right (RMLO)	Left (LMLO)	Left (LCC)	Right (RCC)
C6	Right (RMLO)	Left (LMLO)	Right (RCC)	Left (LCC)

Readers were instructed to take as much time as needed to view each case until they were satisfied with the viewing phase. Readers were informed about the presence of both normal and abnormal cases but no information was provided to them regarding the expected prevalence. Once the reader was prepared to give a diagnostic assessment of the case, the eye-tracking recording process was halted pending completion and reporting of case specific findings, and the reader was ready to proceed with viewing the next case. The readers task was to mark and rate any suspicious findings. Each mark was classified and rated for likelihood of malignancy on a BIRADS-based scale, which consists of five levels (2, 3, 4A, 4B, 4C, and 5) of increasing probability of malignancy [198]. Cases with no markings were assigned a BIRADS rating of 1. After completion of case reporting, the reader was instructed to proceed with the next case. Prior to the every reading session, each reader was carefully calibrated using the 9-point calibration protocol provided by ASL and trained on five training cases selected from the DDSM database. The set of cases used for training were

excluded from the set of cases used in the study.

The cases were presented in a randomized order for each reader using a distinct randomization scheme for each reader. Readers were also permitted to complete the study in multiple sessions based on preference and scheduling conflicts. For example, of the ten readers, two completed case readings for the study in one day (over two sessions), four completed case readings for the study in two days (over at most three sessions), and the remainder completed case readings in three days (over at most 4 sessions).

2.2.3 Data Processing and Feature Extraction

Table 2.4: Enumeration of dual display viewing arrangements and corresponding images on each monitor.

No.	**Dual Display Viewing Arrangement**	**Left Monitor**	**Right Monitor**
1	Same mammographic view (CC)	RCC	LCC
2	Same mammographic view (MLO)	RMLO	LMLO
3	Same breast viewing (Right)	RMLO	RCC
4	Same breast viewing (Left)	LMLO	LCC
5	Four-view (default)	RMLO & LMLO	RCC & LCC

As described in the previous section, gaze data for each reader and each case were collected from four mammographic views spread across two monitors. Raw gaze data was preprocessed using the EyeNAL analysis program from Applied Science Laboratory, which converts raw gaze data to a time-ordered sequence of fixations $f_1, f_2, ..., f_n$, along with other measures associated with fixation (such as fixation duration and inter-fixation degree). These fixations represent a grouping of at least three temporally sequenced raw gaze-position points within 0.5° of visual angle of each other, and a minimum threshold of 100ms total gaze time.

The scanpath, derived by connecting time-ordered fixations or gaze points while viewing each case, resulted in a dense graph representing the visual search process. To measure the complexity of this graph we used the scalar quantity fractal dimension D. Fractal dimension is a mathematical tool for objective measurement of complex structures or patterns which cannot be readily described or quantified through application of Euclidean geometry. The scanpath (visual search) can be treated as a fractal pattern. Its fractal dimension is a non-integer D with the range: $(n - 1 < D \leq n)$ where $n = 2$ is the pattern dimensionality. Using the MinkowskiBouligand box-counting method [108], we estimated D from the scanpath for each case examined by each image reader. Suppose $N(\epsilon)$ is the number of boxes of length ϵ required to cover the scanpath G, we define D_{box} for the two-dimensional graph as:

$$D_{box}(G) \coloneqq \lim_{\epsilon \to \infty} \frac{\log N(\epsilon)}{\log(1/\epsilon)} \tag{2.1}$$

2.2.4 Image Representation and Visual Search

During the reading session, readers typically jump from one of the five possible dual display viewing arrangements (see Table 2.4) to another resulting in a unique non-homogeneous two-dimensional image coordinate space of eye position data for each display view arrangement. To perform fractal analysis of gaze patterns, raw eye position data from each of the unique coordinate spaces was combined to create a single two-dimensional coordinate space, representing eye-position data for each individual case. Table 2.3 enumerates six possible configurations for data representation (i.e., configurations $C_i, i = 1, 2, ..., 6$) for aggregating gaze data into a single coordinate space based on the allowable hanging protocols. The default data representation is the one that corresponds on the default hanging protocol applied at the

beginning of each as illustrated in Figure 2.1a.

We converted raw eye position data for the duration of each case in two steps. First, we mapped gaze position onto a mammographic image-dependent pixel coordinate space to handle zoom, image translation, and other artifacts from eye tracking. Subsequently, each mammographic image, along with respective eye position data were mapped onto a unified pixel coordinate space through a simple translation and scaling (see Equation 2.2).

$$\begin{bmatrix} \acute{x} \\ \acute{y} \\ 1 \end{bmatrix} = \begin{bmatrix} A\cos\theta & -A\sin\theta & d_{xi} \\ A\sin\theta & A\cos\theta & d_{yi} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2.2}$$

where A represents a scaling factor,
θ represents an angle of rotation (set to zero for our purposes), and
d_{xi} and d_{yi} represent translation parameters for the i^{th} mammographic image.

Initial analysis was performed on the data representation corresponding to the default image arrangement (see configuration C6 in Table 2.3). Further, we investigated the effects, if any, of using alternative configurations for data representation (see $C1C5$ in Table 2.3), illustrated in Figure 2.2, on the computed fractal dimension and if any discovered effects alter our initial findings.

2.3 Results

2.3.1 Diagnostic Performance

We grouped each of the 10 participating readers into one of three experience levels: new trainee resident (NR), advanced trainee resident (AR), and expert radiologist (E) as illustrated in Table 2.2. We mapped the diagnostic decision for each case to

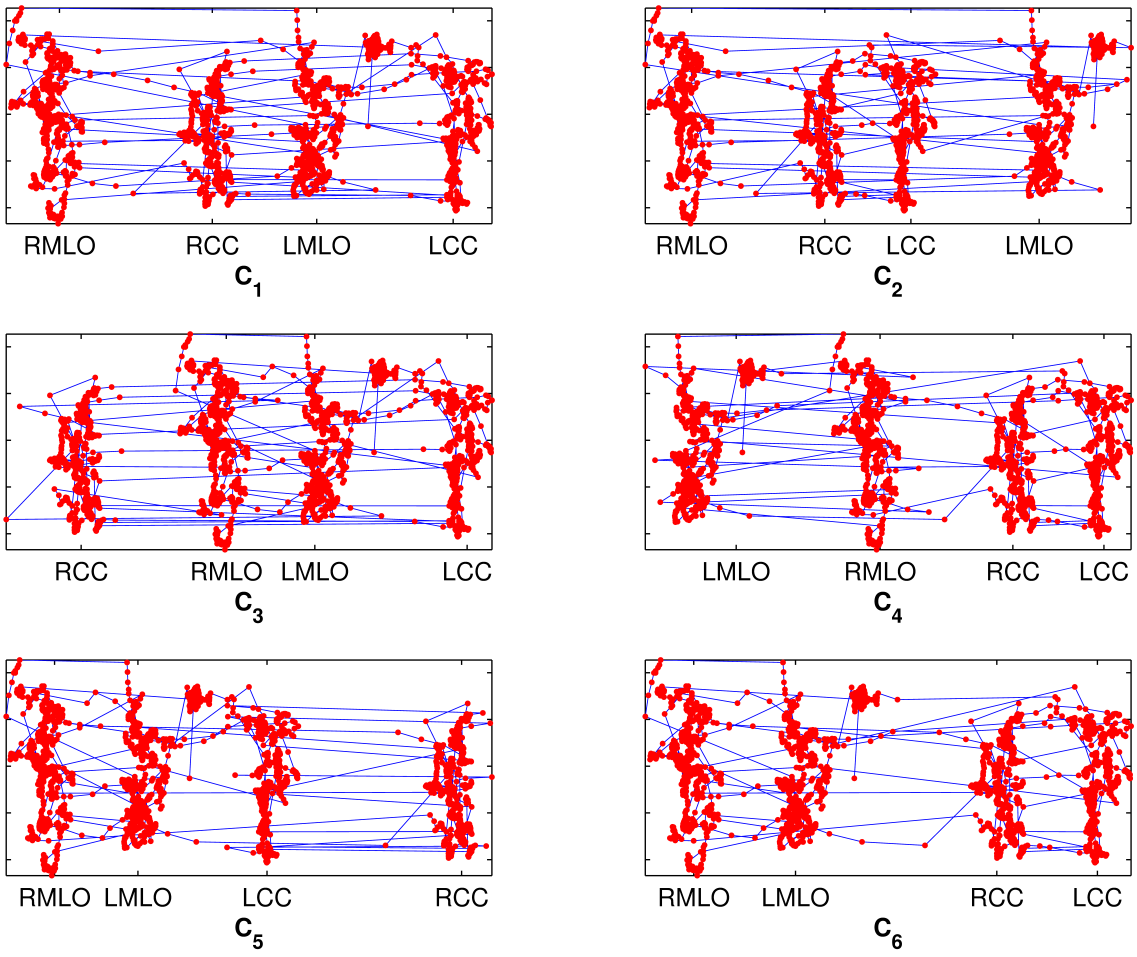


Figure 2.2: Gaze data collected for a single reader synthesized in the 6 possible configurations for data representation.

Table 2.5: Mass detection performance: mass-present (M) vs. mass-absent (N) for new residents (NR), advanced resident (AR), and expert (E) radiologists.

	NR1	**NR2**	**NR3**	**AR1**	**AR2**	**AR3**	**AR4**	**E1**	**E2**	**E3**
True Positive	59	71	62	49	47	49	38	75	72	72
True Negative	12	3	5	18	14	17	17	5	10	9
False Positive	13	22	20	7	11	8	8	20	15	16
False Negative	16	4	13	26	28	26	37	0	3	3
Sensitivity (%)	0.79	0.95	0.83	0.65	0.63	0.65	0.51	1.00	0.96	0.96
Specificity (%)	0.48	0.12	0.20	0.72	0.56	0.68	0.68	0.20	0.40	0.36
Accuracy (%)	0.71	0.74	0.67	0.67	0.61	0.66	0.55	0.80	0.82	0.81

one of the three case pathologies illustrated in Table 2.1 based on the BIRADs rating provided. We designated cases without markings (i.e. no scores were given) as normal (N); we grouped BIRADS ratings 2 and 3 as benign (B); and we grouped BIRADS ratings 4A, 4B, 4C, and 5 as malignant (M). We formed three groupings by breast parenchyma density by combining heterogeneous and dense cases in the same density grouping (due to the small sample size of density 4).

To determine mass detection performance, we compared the BIRADs ratings provided by each reader with the ground truth. We grouped benign and malignant cases under a single class label: mass-present (M), and normal cases under a second class label: mass-absent / normal (N). We report the average diagnostic performance using this two-class grouping (mass-present vs. mass-absent) for each individual radiologist in Table 2.5. From Table 2.5, we computed the average accuracy by experience level: 70.7% $\pm$ 3.5% (new residents), 62.25% $\pm$ 5.5% (advanced residents), and 81% $\pm$ 1.0% (experts). The accuracy of the expert radiologists was significantly higher than that of the advanced residents (two-tailed p-value=0.002) and the new residents (two-tailed p-value=0.008). No significant difference accuracy was observed between new residents and advanced residents (two-tailed p-value=0.077). Readers appeared to execute the clinical task by operating with very different decision criteria in terms of emphasizing sensitivity vs. specificity.

2.3.2 Fractal Dimension of Image Reader's Visual Search

The fractal dimension of image readers scanpath (hereinafter referred to as visual search) ranged between 1.08 and 1.51. In Figure 2.3, we present the average fractal dimension across all cases grouped by case specific properties: case pathology (normal, benign, and malignant), breast density (fatty, fibroglandular, and heterogeneous/dense), and image reader experience level (new Radiology resident, advanced

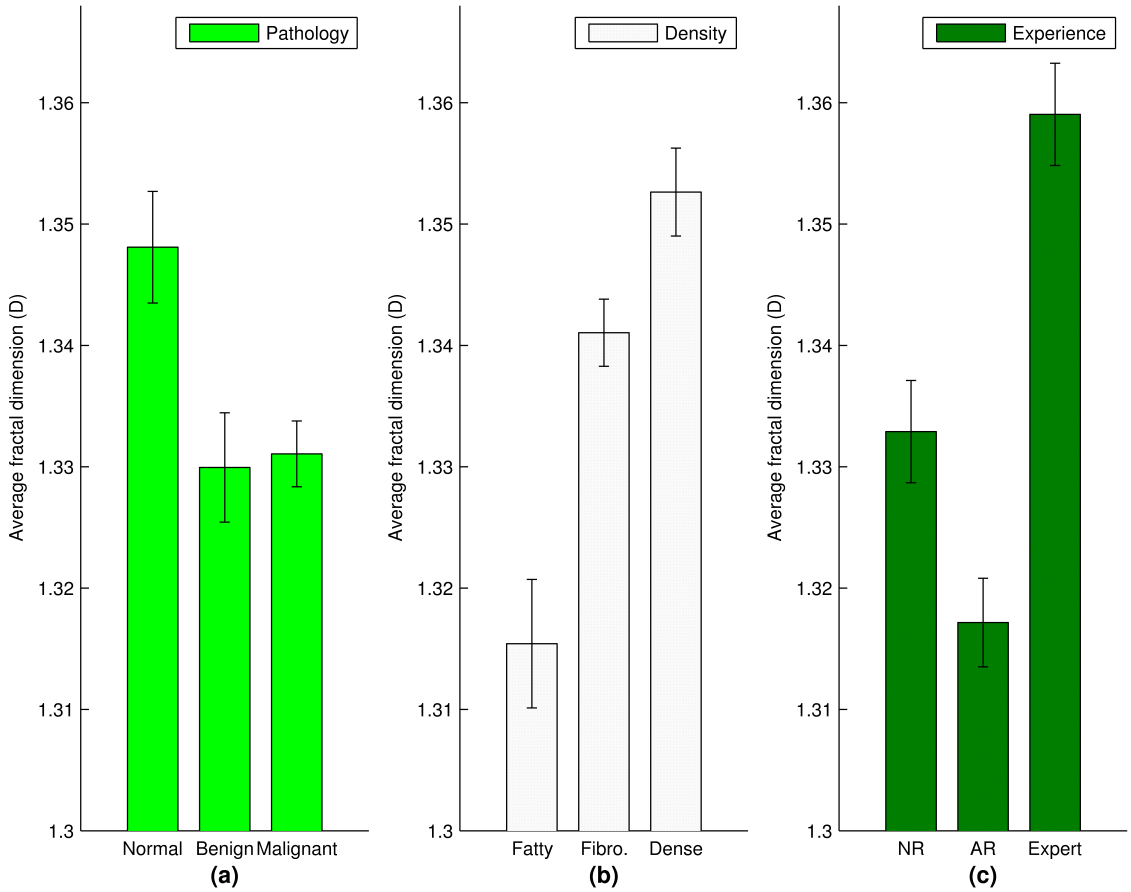


Figure 2.3: Averaged complexity of visual search across case and reader properties: (a) case pathology (normal, benign, and malignant); (b) breast density (fatty, fibroglandular, and heterogeneous/dense); and (c) image reader experience level: new Radiology residents (NR); advanced Radiology residents (AR), and expert radiologists (E).

Radiology resident, and expert radiologist).

2.3.2.1 Effect of Case Pathology on Fractal Dimension of Visual Search

The average fractal dimension for normal cases ($1.35 \pm 4.9e^{-3}$) was significantly higher ($p = 0.01$) than the average fractal dimension for mass-present cases, which contain a benign mass ($1.33 \pm 4.5e^{-3}$), and similarly higher ($p = 0.004$) than the average fractal dimension for mass-present cases, which contain a malignant mass ($1.33 \pm 2.7e{-}3$). However, there were no significant differences ($p = 0.98$) in the average fractal dimension between both types of mass-present cases (malignant . benign).

2.3.2.2 Effect of Mammographic Density on Fractal Dimension of Visual Search

In Figure 2.3b, we observe that the fractal dimension increases monotonically with mammographic density. The average fractal dimension ($1.315 \pm 6e{-}3$) for low-density mammographic cases is significantly lower ($p = 2e{-}4$) compared with the average fractal dimension ($1.34 \pm 3e{-}3$) for medium-density mammographic cases. The average fractal dimension for low-density images was also significantly lower ($p = 8e{-}8$) compared with the average fractal dimension ($1.353 \pm 4e{-}3$) for high-density (heterogeneous/dense) cases. The average fractal dimension for medium-density mammographic cases was also significantly lower than the average fractal dimension for high-density mammographic cases ($p = 2e{-}2$).

2.3.2.3 Effect of Readers Experience Level on Fractal Dimension of Visual Search

Figure 2.3c illustrates the averaged fractal dimension for image readers grouped by experience level. We observed that the average fractal dimension for experienced

radiologists ($1.36 \pm 4e{-}3$) was significantly higher ($p = 3.4e{-}5$) than the average fractal dimension for new Radiology residents ($1.33 \pm 4e{-}3$), and significantly higher ($p = 1e{-}9$) than the average fractal dimension for advanced Radiology residents ($1.32 \pm 3e{-}3$). The average fractal dimension for advanced Radiology residents was significantly lower ($p = 0.01$) than the average fractal dimension for new Radiology residents.

2.3.3 Analysis of Variations in Fractal Dimension of Visual Search

The average fractal dimension of visual search during mammographic screening varied with the characteristics of each case (pathology and density) and with individual image reader as observed in Figure 2.3. Therefore, we performed ANOVA tests on the average fractal dimension computed for each case to determine if there was a dependency with case pathology, breast density, or the image reader's experience level. To analyze the interaction between gaze complexity, case pathology, case density, and image reader experience level, we applied a four-factor fixed-effects ANOVA with three levels for case pathology (normal, benign, and malignant), three levels for breast parenchyma density (fatty, fibroglandular, and heterogeneous/dense), and three levels for experience (new trainee Radiology resident (NR), advanced trainee Radiology resident (AR), and expert radiologist (E)), across 10 individual readers. In Table 2.6, we report the results of ANOVA using average fractal dimension estimated for the six image configurations illustrated in Table 2.3. The ANOVA results showed that all four factors were independent predictors of a radiologists visual search complexity. However, none of the higher order effects were found to be significant. These results were consistent across all six configurations for data representation (see Table 2.3).

Overall, the results from our ANOVA tests show that the pathology and density

Table 2.6: Multi-factor ANOVA test results for possible image configurations

Source	Image Representation					
	C1 $(p > F)$	**C2** $(p > F)$	**C3** $(p > F)$	**C4** $(p > F)$	**C5** $(p > F)$	**C6** $(p > F)$
Pathology	$4e^{-9}$	$2e^{-8}$	$1e^{-7}$	$4e^{-8}$	$4e^{-9}$	$1e^{-8}$
Density	$6e^{-14}$	$7e^{-14}$	$2e^{-13}$	$1e^{-13}$	$8e^{-14}$	$9e^{-15}$
Experience	$2e^{-7}$	$2e^{-5}$	$7e^{-8}$	$2e^{-7}$	$5e^{-7}$	$9e^{-12}$
Individual	$8e^{-60}$	$5e^{-60}$	$1e^{-62}$	$2e^{-65}$	$3e^{-69}$	$2e^{-69}$
Pathology : Density	0.92	0.93	0.91	0.86	0.92	0.88
Pathology : Experience	0.31	0.32	0.21	24	0.29	0.32
Pathology : Individual	0.16	0.1	0.14	0.11	0.7	0.11
Density : Experience	0.62	0.83	0.72	0.78	0.8	0.85
Density : Individual	0.06	0.11	0.11	0.06	0.1	0.03
Pathology : Density : Experience	0.58	0.91	0.87	0.88	0.91	0.48
Pathology : Density : Individual	0.53	0.85	0.77	0.8	0.85	0.32

of a mammographic case both have a significant effect ($p < 1.7e^{-3}$ and $p < 3e^{-6}$ respectively) on visual search complexity as calculated using fractal dimension. The ANOVA tests also show that individual factors (individual differences and level of experience) both have a significant effect ($p < 1e^{-3}$, $p < 9e^{-10}$ respectively) on fractal dimension. These findings indicate that the trends observed in Figure 2.3 and highlighted in Section 2.3.2 are statistically significant. When two or more factors are combined, we find that their combined effects are not statistically significant.

Since results from the ANOVA tests did not depend on the configuration used for data representation, we developed a case-dependent data representation to compute the fractal dimension of the visual search for each case. This approach computed fractal dimension for visual search based on the predominant display arrangement used by the reader for each case. With the thus computed case-dependent fractal dimension of visual search complexity, we applied five-factor fixed-effects ANOVA test by including image reader's diagnostic interpretation as the fifth factor along with pathology, density, experience, and individual differences (as described in Sec-

Table 2.7: Multi-factor ANOVA test results for case based image configurations.

Source	**DF**	**F**	$p > F$
pathology	2	3.91	$2.04e^{-02}$
density	2	11.55	$1.14e^{-05}$
experience	2	6.72	$1.27e^{-03}$
diagnosis	2	4.57	$1.06e^{-02}$
individual	7	22.62	$< 1e^{-15}$
pathology*density	4	3.18	$1.33e^{-02}$
pathology*experience	4	1.44	0.2
pathology*diagnosis	4	0.84	0.5
pathology*individual	14	0.88	0.59
density*experience	4	0.46	0.76
density*diagnosis	4	1.25	0.29
density*individual	14	1.04	0.41
experience*diagnosis	4	1.19	0.31
diagnosis*individual	14	0.77	0.71
pathology*density*experience	8	0.46	0.88
pathology*density*diagnosis	8	0.63	0.75
pathology*density*individual	28	0.43	1
pathology*experience*diagnosis	8	0.93	0.49
pathology*diagnosis*individual	24	0.86	0.66
density*experience*diagnosis	8	1.65	.12
density*diagnosis*individual	26	0.59	.95
pathology*density*diagnosis*experience	11	2.39	$7e^{-3}$
pathology*density*diagnosis*individual	13	1.08	.37
Total	999		

tion 2.3.2). The results from ANOVA tests were consistent with our previous findings. In addition however, the results showed that diagnostic decision was an independent predictor of the fractal dimension of visual search ($p < 0.01$).

Table 2.8: Pairwise comparisons of groups of case pathology, breast density, and radiologists experience level.

Pair 1	**Pair 2**	**p-value**
Pathology Normal	Pathology - Benign	0.01
Pathology Normal	Pathology - Malignant	$4e^{-3}$
Pathology Benign	Pathology - Malignant	.98
Density Fatty	Density - Fibroglandular	$5e^{-3}$
Density - Fatty	Density - Heterogeneous/Dense	$3e^{-3}$
Density - Fibroglandular	Density - Heterogeneous/Dense	$4e{-}3$
Experience - New Resident	Experience - Advanced Resident	0.01
Experience - New Resident	Experience - Expert	$3e^{-5}$
Experience - Advanced Resident	Experience - Expert	$1e{-}9$

Post-ANOVA t-tests with Bonferroni p-value adjustment were also performed and reported in Table 2.8. The fractal dimension of image reader's visual search was significantly different between normal cases and mass-present cases. However, the malignancy status of a mass (benign tumor vs. malignant tumor), did not affect the fractal dimension of visual search. Further, fractal dimension of visual search was found to be significantly different between mammograms of fatty breasts and mammograms of both fibroglandular and heterogeneous/dense breasts. There was also a significant difference in the fractal dimension of visual search between mammograms of fibroglandular breasts and heterogeneous/dense breasts. We also observed that the fractal dimension of visual search was significantly different between all three experience groups: new Radiology residents, advanced Radiology residents, and expert radiologists.

Finally, statistical tests were performed to study the pairwise differences among the gaze networks of the 10 readers (Table 2.9). We noted several significant pairwise differences suggesting that there was substantial inter-reader variability in the fractal dimension of visual search, often among readers of similar experience level.

Table 2.9: Pairwise comparisons of individual readers (new resident resident (NR), advanced resident resident (AR), and expert (E)).

	NR1	**NR2**	**NR3**	**AR1**	**AR2**	**AR3**	**AR4**	**E1**	**E2**
NR2	$< 1e{-}3$								
NR3	1	$< 1e{-}3$							
AR1	$< 1e{-}3$	0.77	0.01						
AR2	.57	$1e{-}3$	1	.27					
AR3	$< 1e{-}3$	.32	$< 1e{-}3$	$1e{-}3$	$< 1e{-}3$				
AR4	.89	$< 1e{-}3$	1	0.08	1	$< 1e{-}3$			
E1	$5e{-}3$	.29	.12	1	.75	$< 1e{-}3$	.39		
E2	.92	$< 1e{-}3$	.34	$< 1e{-}3$	.02	$< 1e{-}3$	.1	$< 1e{-}3$	
E3	$2e{-}3$	$< 1e{-}3$	$< 1e{-}3$	$< 1e{-}3$	$< 1e{-}3$	$< 1e{-}3$	$< 1e{-}3$	$< 1e{-}3$	.22

2.4 Discussion and Conclusions

This study investigated the efficacy of visual gaze complexity for characterizing search behavior of Radiology residents as well as expert radiologists during the diagnostic process for breast cancer in screening mammography. For this study fractal dimension was used as the metric for quantifying the complexity of the visual search patterns. Using a relatively large number of cases, comprising varied pathology and breast parenchyma density, and image readers with varied levels of experience and expertise, the findings presented in this study establish generalizable trends. These trends include the following:

1. The characteristics of a mammographic case (pathology, and breast parenchyma

density) are independent factors in predicting complexity of visual search behavior.

2. The characteristics of the image reader (individual, and level of experience) are independent factors in predicting complexity of visual search behavior.

3. The pathology and breast parenchyma density of a mammographic case, experience level of the image reader, and the resulting diagnostic decision combine as predictors of complexity of visual search behavior during mammographic screening.

4. The visual search complexity while viewing cases with normal pathology are significantly different from cases with malignant pathology.

5. The visual search complexity increases monotonically with increasing breast parenchyma density of a mammographic image. Effectively, low-density mammographic images correspond to lower visual search complexity, while medium-density images correspond to a higher visual search complexity, and high-density images correspond to the highest visual search complexity. This finding is consistent with results obtained by Al Mousa et al. [185], who reported significant increases in visual search parameters when comparing low- and high-density mammograms.

6. On average, the visual search complexity of Radiology residents (both new trainees and advance trainees groups) is significantly lower than the average complexity of experienced radiologists.

7. There are notable differences in visual search complexity between individual radiologists.

This study is novel in its replication of the dual monitor viewing and decision tasks characteristic of screening mammography in clinical practice. It presents a single quantity, fractal dimension, capturing the complexity of visual search behavior during the mammographic screening process. This metric can be leveraged in the future to develop better models for predicting individualized radiologist error risks for a specific case in review. These findings also present future opportunities for personalized decision support and training support technology in Radiology.

Despite the thorough replication of clinical practice, there are notable limitations with this study. While fractal dimension successfully characterizes spatial complexity of visual search, it does not incorporate any temporal information which, intuitively, contain information relevant to readers visual search behavior and diagnostic performance as noted in [143, 141]. We are currently working on developing novel strategies to capture such information. In addition, our study focused specifically on the detection of mammographic masses. It is important to investigate the same issue for other mammographic lesions as well.

Lastly, our study utilized a popular but fairly old dataset of digitized mammograms. A separate study is needed to confirm how our findings would translate in digital mammography. A prior study suggested significant differences in visual scan behavior between screen-film and digital mammograms [176]. However, that earlier study was based on two-view mammograms (single breast viewing). Furthermore, the differences observed in that study involved traditional metrics such as time to first hit and total dwell time. Our study implemented a clinically realistic viewing scenario and a more spatially comprehensive metric of visual search. Furthermore, by providing the full list of the publicly available cases we used we enable other researchers to perform comparative studies.

3. SHAPELET ANALYSIS OF OCULAR CHANGES FOR MODELING VISUO-COGNITIVE BEHAVIOR IN SCREENING MAMMOGRAPHY*

3.1 Introduction

3.1.1 Breast Cancer Screening

Globally, breast cancer is one of the more prevalent forms of cancer within the female population. This form of cancer is the most frequently diagnosed, and the second leading cause of cancer-related deaths among women worldwide [80]. The mortality rate for this cancer is largely dependent on early detection and proper intervention [80]. Through early detection, while the disease is localized, patients have a high (98.5%) relative survival rate. In contrast, survival rates are dramatically lower (25%) when the cancer is metastasized; a point at which the disease becomes incurable [247].

Most patients suffering from breast cancer remain unaware of the disease because it seldom shows any physically visible signs; a common phenomena characteristic of most types of cancer. For this reason, breast cancer is primarily detected and diagnosed through a specialized process known as mammography. Mammography is a medical imaging technique, which uses low-energy X-Rays (approximately $30KVp$) to capture visual representations of the human breasts. These visual representations, known as mammograms, are subsequently examined for the presence of cancer related anomaly. The examination of mammograms, which can be film or digital, is performed by a specialist (board certified radiologists), who is trained in detecting

*Description of methods and experimental results are reprinted with permission from "Shapelet analysis of pupil dilation for modeling visuo-cognitive behavior in screening mammography," by Folami T Alamudun, Hong-Jun Yoon, Tracy Hammond, Kathy Hudson, Garnetta Morin-Ducote, and Georgia Tourassi, 2016. *Proceedings of SPIE*, 9787, pp. 97870M-97870M-13, Copyright 2016 by SPIE.

cancerous anomalies.

There are two types of mammographic examinations performed by radiologists: screening mammography and diagnostic mammography. Screening mammograms are performed to check for the presence of breast cancer in women who are absent of physically visible signs or symptoms of the disease. A screening mammogram generally involves four views from two X-Ray images of each breast: the craniocaudal (CC) view and the mediolateral oblique (MLO) view [119]. These x-ray images make it possible to visually detect tumors, which cannot be detected through physical examination. Screening mammograms also capture microcalcifications (tiny deposits of calcium), which are often an indication that a cancerous growth is also present.

In contrast, a diagnostic mammogram is administered to a patient who has previously demonstrated abnormality in previous clinical inquiry, such as the presence of a lump, a thickening of the skin of the breast, nipple discharge, breast pain, or other physically visible signs on the breast [119]. These abnormalities may be symptoms of some other disease or benign condition however.

3.1.2 Performance in Breast Cancer Screening

The mammographic screening process is not without flaw. Recent studies show the process as being plagued with low sensitivity ($68 - 92\%$ range), with a notably high type II error rate (false-negative) of 29% in visually detectable cancers. Approximately 50% of these diagnostic inaccuracies result from human error. While type I errors (false-positives) at 28.3% can have adverse negative impact/effect on the mental health and well-being of the patient, the occurrence of a type II error has a significant impact on the patients prognosis [67].

Despite the well-studied benefits of screening mammography, the process is not without major drawbacks, which pose significant challenges to the scientific research

community. Most notable among these challenges is the occurrence of error during the mammographic screening process. The *sensitivity* of screening mammography varies because of the multitude of methods used in calculating and reporting [240, 23]. However, the generally accepted range for sensitivity in screening mammography falls between $68 - 92\%$ [297, 138, 225].

Previous studies have shown that the diagnostic interpretation of mammograms is susceptible to different types of human error, which result in missed diagnosis [30, 297, 26]. Studies also showed that human errors resulting from both perceptual and interpretation error, are significant factors responsible for low performance in screening and diagnostic mammography [147, 144, 19]. In an investigation of diagnostic error in screening mammography, Bird et al. analyzed of 320 cancers found in a population of women (average age: 59 ± 0.3), who underwent mammographic screening between 1985 and 1990 [30]. In their study, they categorized missed lesions (false negatives) as cases where: (a) lesion was retrospectively visible; (b) undetected by first reader, but subsequently correctly identified by second reader (double reading); and (c) a cancer correctly diagnosed during a mammography examination immediately preceding the pathologic diagnosis, but retrospectively visible, but incorrectly interpreted as negative, on prior mammograms. They found that 77 cancers were missed at screening mammography for reasons, which include: having a benign appearance; being present at previous screening; only visible in one of the mammographic views of same breast; being located on the site of a previous biopsy; and, in 47% of the cases, from being overlooked [30]. Additional studies conducted over the last few decades have drawn similar conclusions, with the general consensus that human error accounts for nearly 50% of all diagnostic errors [168, 29, 25, 26, 27].

Over time, attempts have been made to improve the sensitivity of diagnostic performance in screening mammography. Notably, the inclusion of complimen-

tary modalities such as ultrasound and magnetic resonance imaging to supplement mammographic information. Independent double reading of screening mammograms, where two radiologists perform readings of a screening mammogram independently to form a consensus, were proposed and adopted. Many studies have shown the effectiveness of this method in increasing the number of detected cancers [8, 266, 41, 66]. However, this process is associated with an increased workload and an associated cost burden (since it presumably entails twice the amount of work).

The two predominant sources of human error in screening mammography are perceptual error and interpretation error [147, 19]; making both sources the subject of a large body of research inquiry. These inquiries are directed primarily to understanding how both errors occur, applying this knowledge in the form of feedback during the diagnostic process, and toward improved training methodology. This class of research focuses on computational understanding of the observer component for the purpose of improving diagnostic performance in mammography [142, 267, 5].

Since diagnostic performance is a measure of the image readers' ability to detect the presence and extent of a cancerous growth, we can infer that factors associated with perception and cognition may also provide quantifiable predictive measures of diagnostic performance. Nodine et al. [193] previously investigated the correlation between number of years of experience, training, and mammography expertise. In this study, Nodine et al. investigated the effect of perceptual and cognitive skills in both detection and interpretation in mammography by analyzing the performance of three categories of image readers (expert radiologists, Radiology residents, and mammography technologists). They found that experts had the highest performance outcomes in comparison with residents, who had a significantly lower performance equivalent to that of mammography technologists. Nodine et al. concluded that poor performance by residents resulted from a lack of perceptual-learning experience dur-

ing mammography training, and recommended a systematic *mentor-guided* training and feedback to improve image perception and decision making [193].

3.1.3 Mental Workload and Task Performance

A goal-oriented search task, similar to the search for a mass within a mammographic image for example, is cognitive in nature [118] and requires the coordination of multiple resources. The mental resource capacity required to perform this search process successfully and efficiently can be measured as the amount of information, which can be maintained and processed in working memory (mental workload or cognitive workload).

Wickens' theory of multiple resources [286] distinguishes between three independent measures of the aforementioned resources. These include:

1. The processing stage, which entails perception of, or output response, to stimulus.

2. The modality of perception, which entails visual or auditory input modalities, or manual or vocal output response.

3. Verbal or spatial codes of perception and central processing.

The construct of the multiple resource theory is consistent with Baddeley's model of working memory [13], which incorporates the concept of separate spatial and verbal components. Baddeley defines the term working memory as a brain system, which provides temporary storage and manipulation of the information necessary for such complex cognitive tasks as language comprehension, learning and reasoning. Working memory requires simultaneous storage and processing of information by its three constituent subcomponents:

1. The attentional-controlling central executive system.

2. The visuospatial sketch pad for manipulating images.

3. The phonological loop for storing and processing auditory information.

Intuitively, mental workload can be conceptualized as characterizing the relationship between task-imposed quantitative demands for resources, and the operator's ability to supply the requisite resources [286]. It characterizes interactions between the processing requirements for completing a task and human resource capacity needed [107]. However, mental resource capacity is limited, varies from one task to another and between individuals [125, 54, 274]. These individual differences in task specific mental resource capacity are consistently found to be correlated with task performance and expertise [286].

Wickens [285] highlighted four workload assessment techniques along with their inherent limitations. These include: primary task measures; secondary task technique; subjective measures; and physiological measures. Of these, the use of psychophysiological measures, which are both unobtrusively and continuously available in real time, individually [20, 1, 167, 190, 3, 228] or in some combination [287, 34], provide a promising path to obtaining measures of mental workload, which correlate with task performance in screening mammography.

3.1.4 Measures of Eye-Movement and Mental Workload

The autonomic nervous system (ANS), a division of the peripheral nervous system, is an unconscious control system, which controls the function of certain bodily organs. It regulates factors such as heart rate, digestion, respiratory rate, pupillary response, arousal etc. in response to external and internal changes. In this regard, the ANS is described as the primary mechanism that controls the fight-or-flight response [121]. Some organs of the ANS, such as the eye (*pupil*), epidermis (eccrine sweat glands), and the cardiovascular organs (various measures of heart rate),

have been investigated as measures of mental workload in task performance prediction [2, 20, 120, 36, 14, 82].

Under conditions of controlled illumination, research has shown pupil dilation as a reliable measure of cognitive and emotional states, including mental workload. Hess and Polt [112] concluded that changes in the size of the pupil during arithmetic multiplication problems can be used as a direct measure of mental workload. They reported an increase of 22% when participants computed a more difficult multiplication problem (16 times 23), compared to an increase of 11% when computing a less challenging problem (7 times 8). Beatty [20] reviewed a large corpus of experimental data and concluded that pupil dilation is a reliable indicator of mental workload, changes in pupil size were positively correlated with changes in mental workload, and that this trend is common across tasks and individuals. However, Beatty [21] also reported differences in individual pupillary response to same task.

Iqbal et al. performed quantitative analysis on the correlation between pupil size and mental workload demand during the execution of interactive tasks, which were designed to represent daily computer-based tasks found in the workplace [120]. Using the average percentage change in pupil diameter, Iqbal et al. were able to differentiate between difficult tasks and less demanding tasks as determined through a subjective ratings of mental workload from NASA Task Load Index (NASA-TLX) survey questionnaire, and task processing times. In a follow-up study, Bailey and Shamsi [14] reported similar findings indicating that mental workload changes throughout the execution of tasks, with pronounced decreases at boundaries indicating task or sub-task initiation and completion. Additional studies have also reported an overall decrease in pupil diameter with increased drowsiness and fatigue during auditory vigilance tasks [154, 303].

3.1.5 Performance Prediction in Screening Mammography

The purpose of the current study is to develop an understanding of the relationship between positional and non-positional measures of changes in the image readers eye during mammographic screening, such as pupil dilation an established measure of mental effort (cognitive workload), the pathological characteristics of a mammographic case, diagnostic decision, and task performance. We focus on examining how mental workload (indexed using measures of pupillary response) during changes the reading of mammographic cases. How these measures vary with both mammographic case pathology, and the image reader's diagnostic performance. To answer these questions, we take advantage of data mining feature extraction algorithms to develop a new index of pupillary response and compare our results with more traditional pupillary response measures. Our ability to answer these questions will provide deeper insight into the effect of mental workload on task performance in screening mammography.

3.2 Materials and Methods

3.2.1 Image Dataset

Table 3.1: Specifications of the 100 four-view screening mammograms used in the study.

Ground Truth	**Patient Age**	**Breast Density**	**Mass Subtlety**	**Total Abnormalities**	**No. of Cases**
Normal	Range: $36-68$ (56.2 ± 10.6)	Range: $1-4$ (Median: 2)	N/A	N/A	25
Benign	Range: $34-82$ (56.9 ± 13.4)	Range: $1-3$ (Median: 2)	Range: $3-5$ (Median: 5)	Range: $1-3$ (Median: 1)	25
Malignant	Range: $37-83$ (64.3 ± 12.4)	Range: $1-4$ (Median: 2)	Range: $1-5$ (Median: 5)	Range: $1-3$ (Median: 1)	50

To perform this study, 100 screen-film mammograms were selected from a corpus of mammographic cases digitized with a high resolution LUMISYS scanner (50m per pixel, 12 bit) from the University of South Floridas Digital Database for Screening Mammography (DDSM) [110]. Each DDSM case contains 4 images, the craniocaudal (CC) and mediolateral oblique (MLO) view images of both the left and the right breasts as well as associated ground truth established via biopsy, additional imaging, or two-year follow-up, radiologists assessment using the BI-RADS^{TM} lexicon [198], and patient age.

Each of the 100 cases was manually selected to cover a broad range of mass margin and shape characteristics. Of the 100 cases selected, 50 included biopsy-proven malignant masses, 25 cases included biopsy-proven benign masses, and the remaining 25 cases were normal as determined during a two-year cancer-free follow-up patient evaluation. Therefore, all mass cases selected for the study included clinically actionable masses. Mammograms with masses deemed as benign-without-callback were excluded. The overwhelming majority of the mass cases (72 out of 75) did not include any microcalcifications. Mass conspicuity was assessed according to the subtlety rating provided in the DDSM truth files. These ratings ranged from 1 (suggesting a subtle lesion) to 5 (suggesting an obvious lesion). A complete list of the DDSM cases used in this study is provided in the Appendix. Table 3.1 provides details on the selected cases, including information on the patients age and breast parenchymal density. The parenchymal density is also provided in the DDSM truth files, and it ranged between 1 (fatty) to 4 (dense), according to the BI-RADS^{TM} lexicon [198].

3.2.2 Experimental Procedure

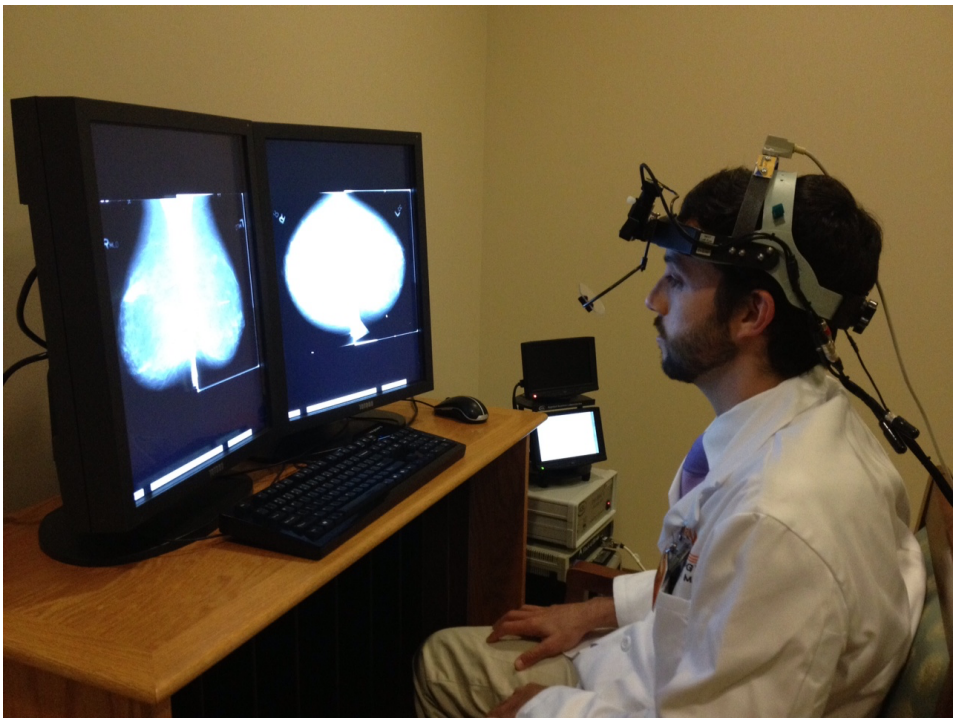
Ten readers of variable experience levels from an academic institution were recruited to conduct blind review of the selected mammograms (see Table 3.2). Each

Table 3.2: Summary of characteristics of study participants.

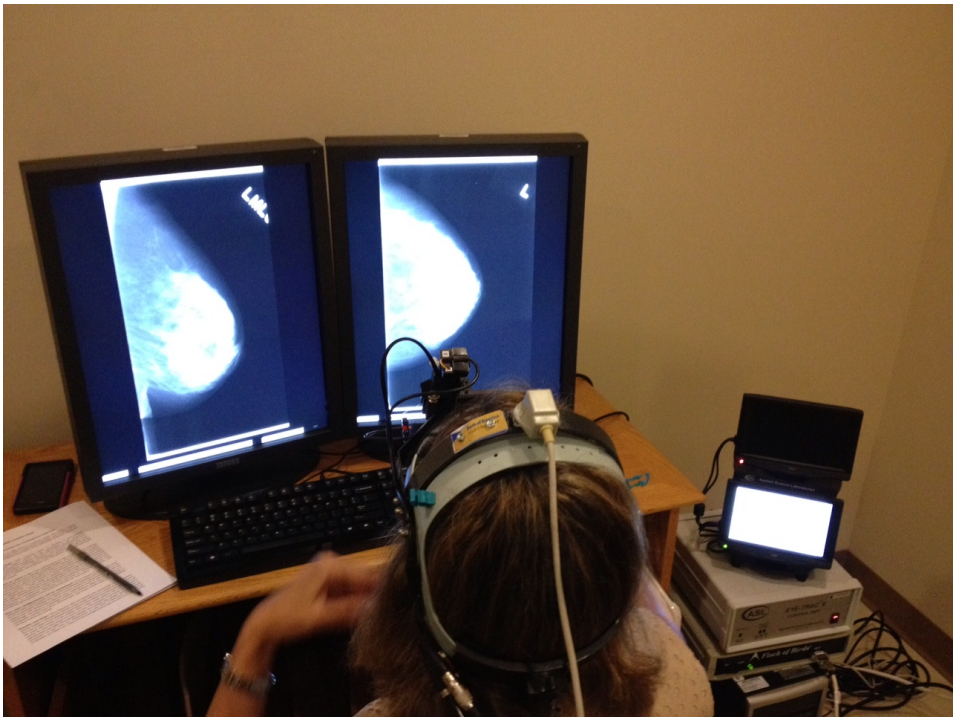
Reader Type	**Experience Level**	**No. of Participants**
Expert Radiologist	$>$ 10 yrs of practice	2
Expert Radiologist	$<$ 10 yrs of practice	1
Advanced Resident	$>$ 2 mammo rotations	4
New Resident	$\leq$ 2 mammo rotations	3
Total		10

reader was asked to report the location of any suspicious mass and provide a corresponding BI-RADS rating as typically done in clinical practice. Of the ten readers, three were experienced MQSA-certified radiologists each with at least nine years of dedicated mammographic experience, four radiology residents with at least three mammography rotations, and three radiology residents with at most two mammography rotations (see Table 3.2). Institutional review board approval was obtained prior to the study. Human subject recruitment and data collection was done according to a protocol approved by the Oak Ridge Site-Wide Internal Review Board. All participants signed an informed consent form.

A customized graphical user interface (GUI) was developed in-house for study participants to view each mammographic case and record their findings. Two medical grade monitors were used (dual-head 5MP mammo-grade Totoku LCD monitors calibrated to the DICOM display standard). The four mammographic views (LCC, RCC, LMLO, RMLO) were initially displayed at low resolution (two views per monitor) to fit the screen. To assess breast symmetry, the users could select the MLO views to be displayed on the left monitor and the CC views to be displayed on the right monitor (e.g., Figure 3.1a). The readers were also able to select and view a single breast at full spatial resolution with the MLO view displayed on the left monitor and the CC view displayed on the right monitor (e.g., Figure 3.1b). Table 3.3



(a) Dual Display showing default arrangement.



(b) Dual display showing MLO and CC views of the right breast.

Figure 3.1: Image reader outfitted with eye-tracking apparatus reviewing a mammographic case.

enumerates all possible "hanging protocols" implemented in the GUI. Please note that based on the allowable protocols RMLO could never appear on the right monitor while LCC could never appear on the left monitor. In addition, the GUI provided the functionality of zooming in/out, panning, and magnifying glass for detailed reading of each mammographic view. During the reading sessions, each reader was outfitted with an H6 head-mounted eye-tracker, with a 60 Hz sampling rate, and eye-head integration from Applied Science Laboratories (ASL, Bedford, Massachusetts, USA). The eye-tracker recorded each readers eye position data to within 0.5° of accuracy.

Table 3.3: Possible configurations for a combined two-dimensional data representation.

	Left Monitor		**Right Monitor**	
Label	**Left Image**	**Right Image**	**Left Image**	**Right Image**
C1	Right (RMLO)	Right (RCC)	Left (LMLO)	Left (LCC)
C2	Right (RMLO)	Right (RCC)	Left (LCC)	Left (LMLO)
C3	Right (RCC)	Right (RMLO)	Left (LMLO)	Left (LCC)
C4	Left (LMLO)	Right (RMLO)	Right (RCC)	Left (LCC)
C5	Right (RMLO)	Left (LMLO)	Left (LCC)	Right (RCC)
C6	Right (RMLO)	Left (LMLO)	Right (RCC)	Left (LCC)

Readers were instructed to take as much time as needed to view each case until they were satisfied with the viewing phase. Readers were informed about the presence of both normal and abnormal cases but no information was provided to them regarding the expected prevalence. Once the reader was prepared to give a diagnostic assessment of the case, the eye-tracking recording process was halted pending completion and reporting of case specific findings, and the reader was ready to proceed with viewing the next case. The readers task was to mark and rate any suspicious findings. Each mark was classified and rated for likelihood of malignancy on a BIRADS-based

scale, which consists of five levels (2, 3, 4A, 4B, 4C, and 5) of increasing probability of malignancy [198]. Cases with no markings were assigned a BIRADS rating of 1. After completion of case reporting, the reader was instructed to proceed with the next case. Prior to the every reading session, each reader was carefully calibrated using the 9-point calibration protocol provided by ASL and trained on five training cases selected from the DDSM database. The set of cases used for training were excluded from the set of cases used in the study.

The cases were presented in a randomized order for each reader using a distinct randomization scheme for each reader. Readers were also permitted to complete the study in multiple sessions based on preference and scheduling conflicts. For example, of the ten readers, two completed case readings for the study in one day (over two sessions), four completed case readings for the study in two days (over at most three sessions), and the remainder completed case readings in three days (over at most four sessions).

3.2.3 Data Pre-Processing

Pupil dilation and eye movement data were monitored and recorded using an H6 head-mounted eye-tracker, with a 60 Hz sampling rate, and eye-head integration from Applied Science Laboratories (ASL). The participants were free to spend as much time reviewing each case. Participants were also free to halt the experiment at the completion of any case *n*, to resume the next case *n+1* at a later time or date based on convenience. Since the time duration for review varied by case and by radiologist, the time window during which the signal of interest was extracted was calculated based on start and stop log entry events. Pupil diameter, fixations, saccade length, and raw eye position related data were made available through accompanying ASL software. The sensor measurement representing the pupil diameter for each case

recording was first linearly interpolated to account for eye-blinks, then downsampled by a factor of ten, and subsequently filtered using an adaptive Hampel [213] filter to remove noise artifacts resulting from equipment induced noise, drift, eye tremors, an other sources.

3.2.4 Measurements and Feature Extraction

3.2.4.1 Pupillary Response Measures

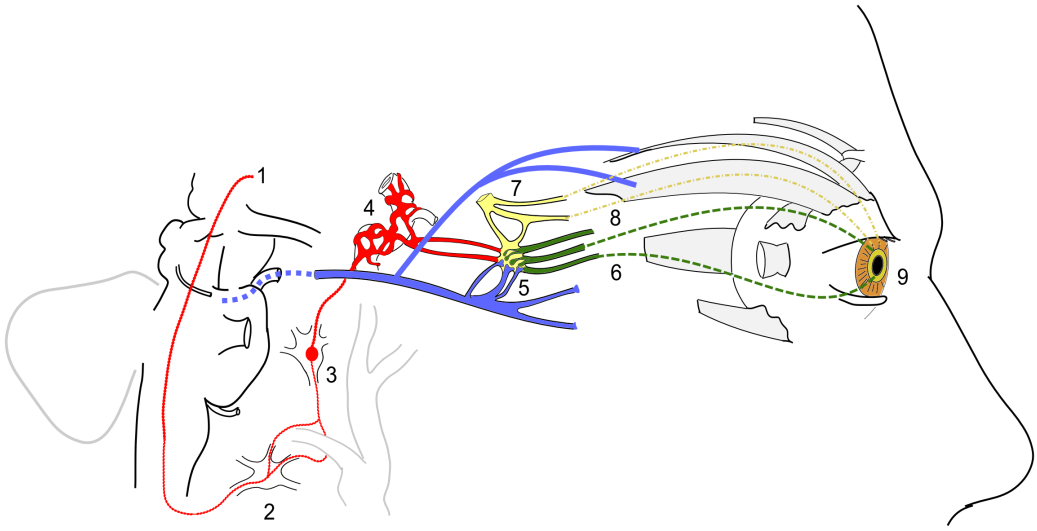


Figure 3.2: Sympathetic and parasympathetic innervation of the pupil. (1) sympathetic fibers arise from the hypothalamus, (2) the stellate ganglion, (3) synapse at the superior cervical ganglion, (4) sympathetic plexus around internal carotid artery, (5) oculomotor nerve (Cranial nerve 3) fibers synapse at the ciliary ganglion (blue), (6) Short ciliary nerves from ciliary ganglion carrying parasympathetic supply to sphincter pupillae (green), (7) Trigeminal fibers (Cranial nerve 5) relay in ciliary ganglion and carry sympathetic supply (yellow), (8) Long ciliary nerve fibers (from the ophthalmic branch of cranial nerve 5) carrying sympathetic supply to the dilator pupillae, (9) Sphincter pupillae (circular fibers) and Dilator pupillae (radial fibers) muscles of the pupil. (From Rajan [227])

The structure of the eye can be compared to a camera: containing a lens for light refraction and focus, and an adjustable aperture for controlling the amount of light admitted. This aperture, the pupil, undergoes large changes, which are controlled by the parasympathetic nervous system (PNS) through the circular muscle (constriction), and the sympathetic nervous system (SNS) through the radial muscle (dilation) 3.2. However, internal physiological responses within the body are also reflected as minute changes in the size of the pupil. These small-scale, rapid changes in the diameter of the pupil are reflective of the dynamic changes in the central nervous system (CNS) that underlie human cognition (see Figure 3.3). Historically, these changes are indexed by computing the percentage change in pupil dilation (ΔP) from a baseline pupil measurement (P_b). P_b was computed as the average pupil size for the first 500ms of pupil recording after initial light reflex.

3.2.4.2 Time Series Shapelet

Typical classification techniques used in machine learning are not efficient at handling real-valued ordered time series data, where temporal ordering and trends of data are as useful in providing discriminative information as discrete values. Such changes are difficult and often impossible to capture using well-known methods such as nearest neighbor algorithm, where features are individual data points and temporal characteristics are not preserved. These challenges result in the need for a type of data primitive, which (a) capture temporal changes of observed data, (b) generate temporal attributes effective for establishing sufficient criteria for class membership and thus usable as a feature for classification, and (c) can be utilized during post-analyzed for model characteristics responsible for determining class membership.

One such concept was introduced recently by Ye and Keogh [299]. In their work on data mining images in historical documents, Ye and Keogh [299] developed a

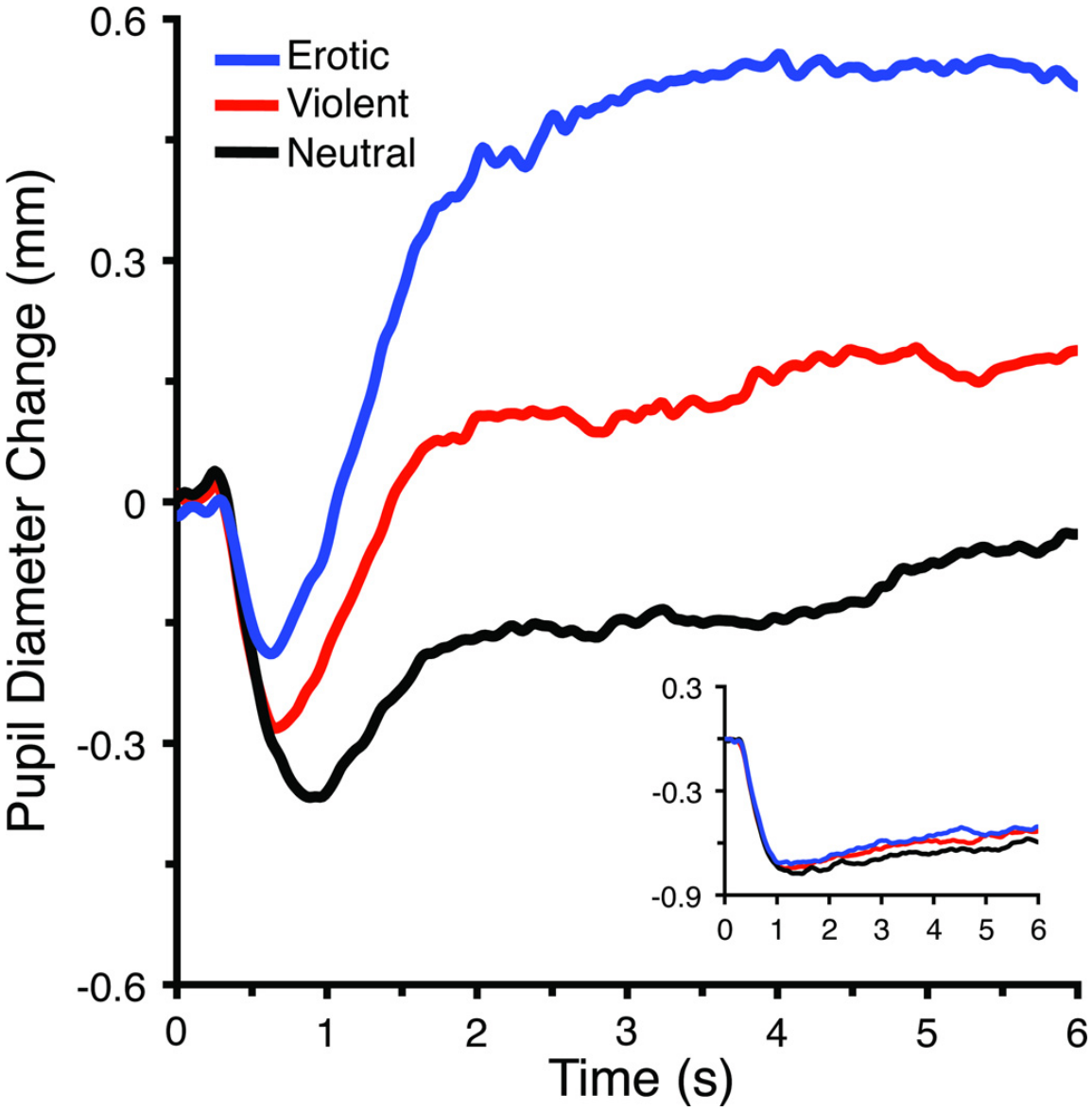


Figure 3.3: Change (mm) in pupil diameter from a 1s baseline preceding picture onset when viewing erotic, neutral, or violent scenes. Inset: For scrambled pictures, the light reflex did not differ as a function of original picture content. (Reprinted with permission from Henderson et al. [111])

novel temporal primitive called *time series shapelets*. They define shaplets as time series sub-sequences, which are maximally representative of a class. We can think of shapelets as features that capture temporal changes in data, which can improve a model's ability to discern between two or more classes.

To illustrate the concept of time series shapelets, Ye and Keogh considered a two-class classification problem of identifying two commonly confused plants, *Uritica dioica* (often called common nettle or stinging nettle) and *Verbena uriticifolia* (white vervain). Converting each sample into a one-dimensional representation, creates a data representation that can be used for classification using one of many existing techniques. Traditional techniques for encoding discriminative information are known to perform poorly on this type of problem [299], primarily because the differences between the two classes are captured in the temporal changes in their respective data representation.However, a shapelet subsequence allows us to compare temporal characteristics of both classes to successfully discriminate between the two. Using Ye and Keogh's methods [299], we use timeseries shapelet analysis to characterize temporal properties in image readers' pupil dilation during a mammographic screening.

3.2.4.3 Measures of Eye Movement

The movement of the human eye is controlled by three pairs of muscles. The combined and coordinated actions of these muscles (depicted in Figure 3.4) are responsible for horizontal (yaw), vertical (pitch), and torsional (roll) eye movements, respectively, and hence control the three-dimensional orientation of the eye inside the head. According to Donders law [273], orientation uniquely decides the direction of gaze, independent of how the eye was previously orientated. Large sections of the brain control these muscles to direct the gaze to the desired locations in space.

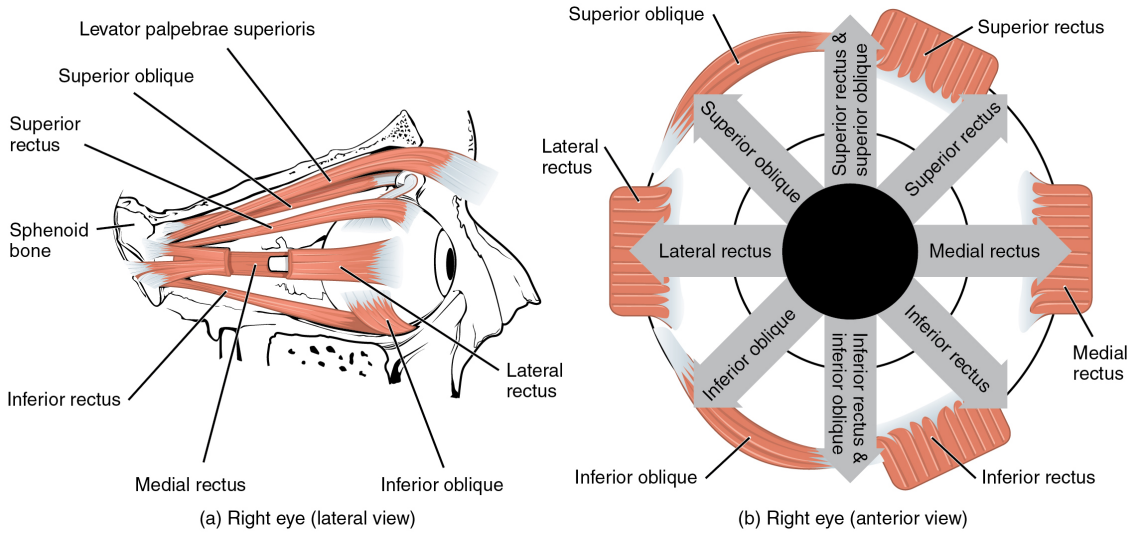


Figure 3.4: An illustration of the human eye muscles that generate the vertical up-down movements (superior and inferior rectus), the horizontal outward-inward movements (lateral and medial rectus), and the torsional rotating movement (superior and inferior oblique)(From OpenStax College, [52]).

Humans and other primates (including other vertebrates) primarily engage in seven types of voluntary and involuntary eye movement: fixation, saccade, glissade, smooth pursuit, microsaccade, tremor, and drift (see Table 3.4) [115]. We recorded gaze data for each image reader from the four mammographic views (RMLO, LMLO, RCC, LCC) spread across two monitors (for each mammographic case). First, we processed raw gaze data to extract fixations. A *fixation* refers to a state where the eyes remain still (within a small radius) over a period of time, such as is the case when the eyes pause on a given word while reading. Fixating on a point or region is generally considered as a measure of attention to a given position or region of interest, even though this is not always the case.

While there is no universally excepted method for detecting fixations, there are established parameters based on ocular physiology, which permit a reasonable criteria for detecting and extracting fixations from gaze data. A typical algorithm to

Table 3.4: Basic measures of positional eye movement events.

Description	**Duration (ms)**	**Amplitude**	**Velocity**
Fixation	$200 - 300$	N/A	N/A
Saccade	$30 - 80$	$4 - 20°$	$30 - 500°/s$
Glissade	$10 - 40$	$0.5 - 2$	$20 - 140°/s$
Smooth Pursuit	N/A	N/A	$10 - 30°/s$
Microsaccade	$10 - 30$	$10 - 40'$	$15 - 50°/s$
Tremor	N/A	$< 1'$	$20'/s$ (peak)
Drift	$200 - 1000$	$1 - 60'$	$6 - 25'/s$

determine a fixation event uses the mean X and Y eye position coordinates measured over a minimum period of time during which the eye does not move more than some maximum amount. This algorithm requires that a point-of-gaze must continuously remain within a small area (approximately within 1° visual angle in our algorithm) for some minimum amount of time (approximately 100ms for our algorithm). From this measure we computed the following features: rate of fixation (F_R), and the average duration of fixations (F_D) on a per case basis.

The eye is not completely still during a fixation, but exhibits three distinct types of micro-movements: *tremor*, *microsaccades* and *drifts* [169]. A tremor is a small movement of approximately 90 Hz. The exact role of tremors is still a subject of research; it is generally believed to be imprecise muscle control. Drifts are slow movements that shift the eye away from the centre of fixation, while the counter movement, a microsaccades, serves to quickly return the eye back to the center of fixation. However, these smaller, faster movements were not computed or utilized in this study.

The rapid motion of the eye from one fixation to another, from word to word while reading, for instance, is called a *saccade*. Saccades are considered the fastest movement the body can produce; typically taking $3080ms$ to complete. It is a gen-

erally held view that human beings are perceptively blind during most of a saccadic event.

An important characteristic of saccades is that they rarely take the shortest path between two points, but instead undergo one of several *shapes* and *curvatures*. Since a saccade is described in terms of the gaze data between detected fixations, we computed saccadic events as gaze points connecting the completion of one fixation to the beginning of the next fixation. From the saccadic measure, we computed the following feature: saccadic amplitude (S_A). The *scanpath* is described as the eye-movement pattern that describes the route of occulomotor events through space within a defined timespan (such as the duration of a mammographic reading). From this measure, we compute the following feature: length of scanpath (SP_L).

The saccadic movement is not mechanically precise, that is they do not stop directly at the intended target, but instead the eye wobbles before coming to a stop. This post-saccadic movement is referred to as glissadic movements or *glissade*. The movement of the eye is characteristically different in the case of following or tracking a moving object such as a bird flying across the sky. This type of eye movement is usually slower and referred to as *smooth pursuit*. The difference between the saccadic and the smooth pursuit movements is that the latter is driven by and requires a moving target, while the former can be made independent of any visual stimulus. This type of movement isn't evidenced in our experiment and was neither computed nor utilized.

3.3 Analysis and Results

3.3.1 Statistical Pattern Analysis

As preliminary analyses, we examined the eye activity responses recorded during the experiment. We grouped each of the twn participating readers into one of three

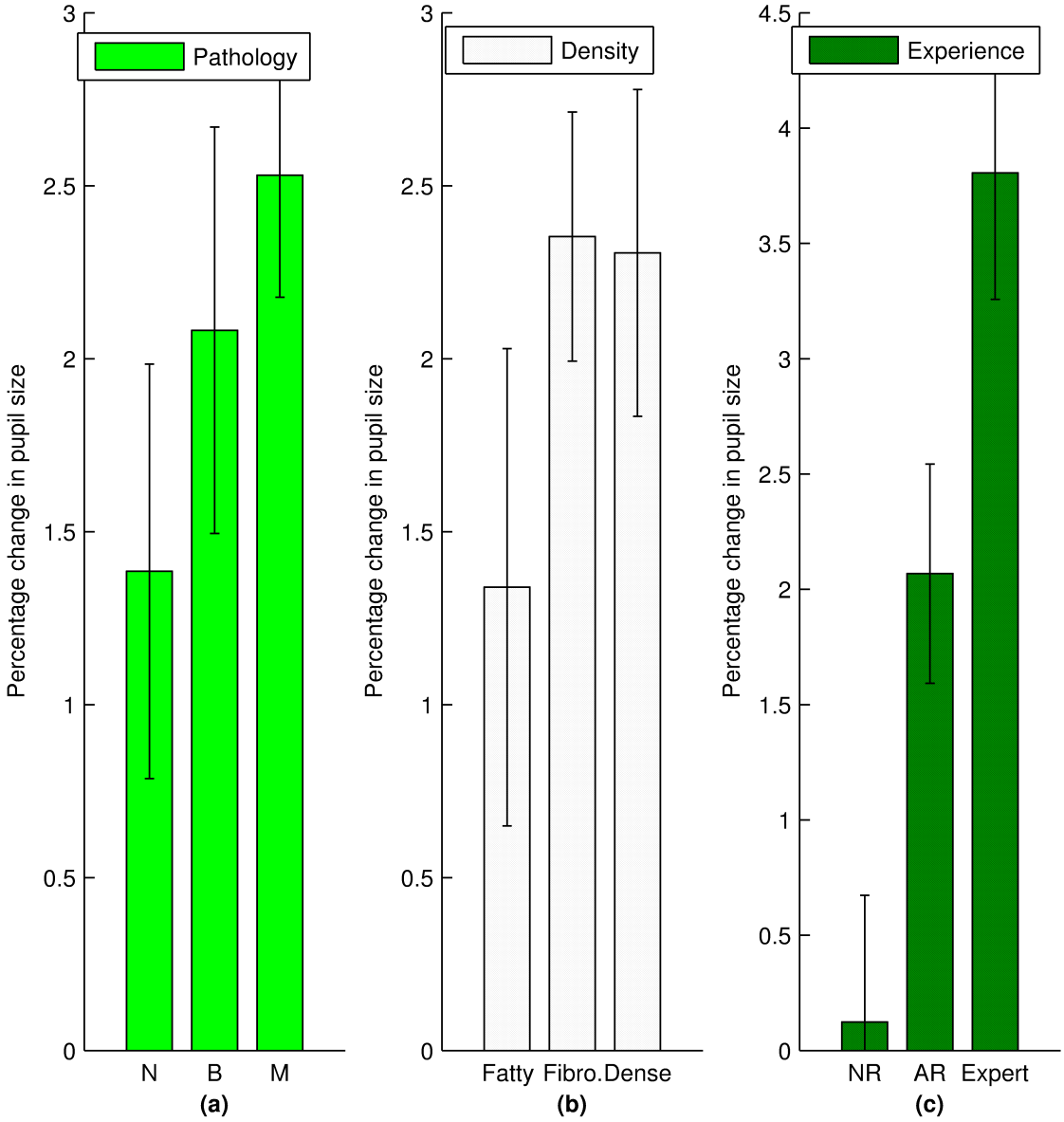


Figure 3.5: Percentage change in pupil dilation. (a) case pathology (normal, benign, and malignant); (b) breast density (fatty, fibroglandular, and heterogeneous/dense); and (c) image reader experience level: new Radiology residents (NR); advanced Radiology residents (AR), and expert radiologists (Expert).

experience levels: new trainee resident (NR), advanced trainee resident (AR), and expert radiologist (E) as illustrated in Table 3.2. We mapped the diagnostic decision for each case to one of the three case pathologies illustrated in Table 3.1 based on the BIRADS rating provided. We designated cases without markings (i.e. no scores were given) as normal (N); we grouped BIRADS ratings 2 and 3 as benign (B); and we grouped BIRADS ratings 4A, 4B, 4C, and 5 as malignant (M). We formed three groupings by breast parenchyma density by combining heterogeneous and dense cases in the same density grouping (due to the small sample size of density 4).

In Figure 3.5 we present an index of changes in pupillary response (average percentage change in pupil diameter) across all cases grouped by case specific properties: case pathology (normal, benign, and malignant), breast density (fatty, fibroglandular, and heterogeneous/dense), and readers experience level (new Radiology resident, advanced Radiology resident, and expert radiologist).

Table 3.5: Summary of eye movement features.

Feature	Pathology			Density			Experience		
	Normal	Benign	Malignant	Fatty	Fibro.	Dense	New R.	Adv. R.	Expert
Pupil Dilation	1.5 ± 0.6	2.5 ± 0.6	2.75 ± 0.4	1.5 ± 0.7	2.8 ± 0.4	2.5 ± 0.5	0.8 ± 0.7	2.4 ± 0.4	4.2 ± 0.6
Pupil Size	32.9 ± 0.4	32.9 ± 0.4	32.8 ± 0.3	32.6 ± 0.5	32.9 ± 0.3	33.2 ± 0.3	36.8 ± 0.5	34.3 ± 0.3	27.7 ± 0.4
No. Fixations	118 ± 6	99 ± 6	97 ± 3	83 ± 7	112 ± 4	120 ± 5	82 ± 6	107 ± 4	125 ± 5
Fixation Rate	2.7 ± 0.1	2.6 ± 0.1	2.5	2.6 ± 0.1	2.5	2.6	2.6 ± 0.1	2.8	2.5 ± 0.1
Fixation Duration	230 ± 3	220 ± 3	230 ± 2	220 ± 4	230 ± 2	230 ± 3	240 ± 4	240 ± 2	200 ± 3
Saccade Amplitude	5.5 ± 0.1	5.5 ± 0.1	5.6 ± 0.1	5.8 ± 0.2	5.5 ± 0.1	5.4 ± 0.1	5.8 ± 0.1	5.2 ± 0.1	5.5 ± 0.1
Saccade Duration	102 ± 4	106 ± 4	106 ± 2	108 ± 5	105 ± 3	101 ± 3	90 ± 4	79 ± 3	144 ± 4
Scanpath (1k pixels)	17.7 ± 0.4	17.3 ± 0.4	17.8 ± 0.2	18.4 ± 0.5	17.2 ± 0.2	17.2 ± 0.3	22.5 ± 0.4	17.2 ± 0.3	13.1 ± 0.4

3.3.1.1 Effect of Case Pathology on Change in Pupil Dilation

In Figure 3.5a, we observe that the percentage change in pupil dilation (*pcpd*) increases monotonically with the BIRADS rating of the mammographic cases. The lowest averages *pcpd* were recorded when readers viewed normal cases ($\mu = 1.54, \sigma = 0.64$), while the highest averages *pcpd* were recorded when readers viewed cases with benign or malignant status (benign cases: $\mu = 2.47, \sigma = 0.63$; malignant cases: $\mu = 2.74, \sigma = 0.38$). An independent-samples t-test was conducted to compare the differences in recorded *pcpd* average between the three case pathology pairings. We observed a significant difference between the average *pcpd* for normal and malignant cases ($t(748) = -1.97, p = 0.05$). We observe no significant differences between the normal and benign ($t(498) = -1.68, p = 0.09$), and the benign and malignant ($t(748) = -0.01, p = 0.99$) case pairings. These results suggest that the underlying pathology of a mammographic case dictates the average percentage change in pupil dilation during mammographic screening. This difference is more pronounced between cases of normal and malignant pathology. Specifically, our results suggest that *cognitive processing* by an image reader (indexed using percentage change in pupil dilation) is significantly higher during mammographic the reading of malignant cases, compared with *cognitive processing* during the reading of normal cases.

3.3.1.2 Effect of Mammographic Density on Change in Pupil Dilation

Figure 3.5b illustrates the average percentage change in pupil dilation grouped by mammographic density. We observe that the average *pcpd* for low-density mammographic cases (fatty cases: $\mu = 1.48, \sigma = 0.73$) is lower when compared with the average *pcpd* for medium-density mammographic cases (fibroglandular cases: $\mu = 2.75, \sigma = 0.38$); heterogeneous/dense cases: $\mu = 2.53, \sigma = 0.5$). However, differences in average *pcpd* between low-density and higher density cases are not

statistically significant (fatty and fibroglandular: $t(718) = -1.75, p = 0.08$; fatty and heterogeneous/dense: $t(528) = -1.38, p = 0.17$). Differences between higher density (medium-density and high-density) cases were not statistically significant (fibroglandular and heterogeneous/dense: $t(748) = 0.23, p = 0.82$).

3.3.1.3 Effect of Experience Level on Change in Pupil Dilation

Figure 3.5c illustrates the average percentage change in pupil dilation for image readers grouped by experience level. We observe that the average *pcpd* increases monotonically with image readers' experience level. The average *pcpd* recorded for expert radiologists ($\mu = 4.24, \sigma = 0.55$) was significantly higher ($t(498) = -5.18, p = 3e-7$) than the recorded averages for new Radiology residents ($\mu = 0.08, \sigma = 0.67$), and similarly higher ($t(798) = 2.77, p = 0.006$) than more advanced Radiology residents ($\mu = 2.44, \sigma = 0.43$). We also observe that the average *pcpd* for advanced Radiology residents was significantly higher than new Radiology residents ($t(698) = 3.44, p = 6e-4$).

These results suggest expert radiologists demonstrate higher *cognitive processing* during mammographic reading, when compared with Radiology residents. Our results also suggest that more experienced Radiology residents demonstrate higher *cognitive processing* than do newer, less experienced Radiology residents.

Interestingly, when we analyzed pupil size (normalized for individual differences), which is an index for cognitive load (how difficult the task is), we observe that the average pupil size decreases monotonically with image readers' experience level. This trend suggests that expert radiologists, experiencing lower *cognitive demands*, find the task of mammographic screening easier, when compared with Radiology residents, who exhibit higher levels of *cognitive demand*.

3.3.2 Predictive Models Utilizing Aggregate Measures of Eye Events

To assess the predictive performance of features derived from measures of eye event, we created three test cases. The first test case was to ascertain the efficacy of aggregate measures of eye events to predict the pathology of a mammographic case during screening. That is, predict what the image reader is looking at (ground truth pathology) using eye movement features during screening. The second case tested the efficacy of aggregate measures of eye events to predict the image readers' interpretation of the mammographic case. That is, predict the image readers' diagnostic interpretation of the case. The third test case examines the efficacy of aggregate measures of eye events to predict diagnostic accuracy during screening. That is, predict image reader performance.

We conducted performance tests on the eye movement feature set by developing a within-subject predictive model using a Random Forest classifier for each test case [37]. All predictive models were evaluated using a k-fold cross-validation scheme ($k = 10$). K-fold cross-validation involved partitioning the data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). Multiple (k) rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds in order to reduce variability. The aggregated predictive values over all rounds serves as the final performance evaluation of the predictive model. All training and testing evaluations were performed using WEKA software package [84], an open source machine learning software for building and testing predictive models.

In Table 3.6, we present results for area under receiver operating characteristics (ROC) curve (AUC) on predicting ground-truth pathology, readers' diagnostic inter-

pretation, and readers' diagnostic performance using eye movement features for each image reader. Although the distribution of *correct* versus *incorrect* diagnoses varied among image readers, each data set provided sufficient samples from each class to allow modeling the risk of diagnostic error for all study participants individually.

Table 3.6: Results for predicting ground-truth pathology, reader interpretation, and reader performance using eye movement features (F_{eye}).

		Ground Truth		***Diagnosis***		***Performance***	
ReaderID	**Experience**	F_{eye}	**ZeroR**	F_{eye}	**ZeroR**	F_{eye}	**ZeroR**
001	Adv. Resident	0.64	0.47	0.69	0.46	0.62	0.46
002	Expert	0.59	0.47	0.59	0.45	0.57	0.47
003	New Resident	0.51	0.47	0.62	0.45	0.59	0.46
004	Expert	0.47	0.47	0.43	0.47	0.57	0.46
005	Adv. Resident	0.4	0.47	0.57	0.47	0.51	0.46
006	Adv. Resident	0.58	0.47	0.75	0.46	0.53	0.48
007	Expert	0.65	0.47	0.6	0.47	0.34	0.45
008	New Resident	0.55	0.47	0.53	0.39	0.5	0.5
009	New Resident	0.57	0.47	0.57	0.44	0.55	0.45
010	Adv. Resident	0.46	0.47	0.76	0.47	0.59	0.45
Average		0.54	0.47	0.61	0.45	0.54	0.46

3.3.3 Predictive Models from Time Series Shapelets

To further investigate the statistically significant nature of the effect of eye movement features on mammographic case properties, we performed within-subject time series shapelet analysis on percentage change in pupil dilation to improve on predictive performance. First, we developed a dictionary of 50 maximally informative shapelets of varied lengths ($1s - 3s$) using the methods presented in [299]. Next, using each unique shapelet as a search term, we computed the *term frequencyinverse document frequency* [156, 249] ($tf - idf$) score, a commonly used term weighting scheme in information retrieval systems, for each shapelet in the dictionary. Using

Table 3.7: Results for predicting ground-truth pathology, reader interpretation, and reader performance using time series shapelets from percentage change in pupil size.

		Ground Truth		***Diagnosis***		***Performance***	
ReaderID	**Experience**	**Shapelets**	**ZeroR**	**Shapelets**	**ZeroR**	**Shapelets**	**ZeroR**
001	Adv. Resident	0.81	0.47	0.79	0.46	0.88	0.46
002	Expert	0.81	0.47	0.81	0.45	0.91	0.47
003	New Resident	0.91	0.47	0.87	0.45	0.86	0.46
004	Expert	0.81	0.47	0.78	0.47	0.89	0.46
005	Adv. Resident	0.85	0.47	0.85	0.47	0.86	0.46
006	Adv. Resident	0.84	0.47	0.82	0.46	0.84	0.48
007	Expert	0.85	0.47	0.79	0.47	0.9	0.45
008	New Resident	0.84	0.47	0.76	0.39	0.84	0.5
009	New Resident	0.83	0.47	0.73	0.44	0.89	0.45
010	Adv. Resident	0.86	0.47	0.76	0.47	0.89	0.45
Average		0.84	0.50	0.80	0.45	0.87	0.46

this method, we represent each mammographic case reading as a vector of $tf - idf$ scores of length 50. representing each shapelet (feature). This method of representation using vectors in a common vector space is known as the vector space model and is fundamental to a host of information retrieval operations ranging from scoring documents on a query, document classification and document clustering [237].

Because the number of features was comparatively high (50) in proportion to the number of data samples (mammographic cases) per image reader (100), along with the associated increase in memory and computation costs, we performed dimensionality reduction (feature subset selection) for each image reader's data set. We used a wrapper feature subset evaluation method [137], which searches for an optimal subset of 10 or fewer features by evaluating feature subsets using a learning scheme. The search function was performed using the simple genetic algorithm described by Goldberg in [78]. The accuracy of the learning scheme for each subset of features was estimated using a Random Forest [37] classifier with 10-fold cross-validation. The final set of features selected for use in the optimal subset were those features, which were selected in at least half of the ten folds in the wrapper subset evalua-

tion. All training and testing evaluations were performed using the WEKA software package [84].

3.3.4 Classification Results

For the performance comparison purposes, we include the results of a ZeroR classifier. The ZeroR classifier is a simple majority rule or mode rule classifier, which always returns the majority or modal class for all input test samples independent of the feature values of the input sample. The results of the ZeroR classifier is equivalent to random chance and serves as baseline for both feature-based and shapelet-based classifiers.

3.3.4.1 Predicting Ground-Truth Case Pathology

In this test, we examine the accuracy with which the image reader's eye movement features predict the pathology of the case being reviewed. On average, the feature-based classifier predicted the pathology of the case being viewed with 54% accuracy. The performance of the feature-based classifier was significantly higher ($p = .02$) than the baseline classifier (47%). However, the average performance results from the shapelet-based classifier (84%) was significantly higher ($p = 4e - 6$) than the results from the feature-based classifier, and significantly higher ($p = 2e{-}11$) than the results from the baseline classifier. These results, illustrated in Figure 3.6, are congruent with previous findings, which indicate that pupillary responses are sensitive to the nature of a given task.

3.3.4.2 Predicting Image Reader's Diagnostic Interpretation

This test examines the accuracy with which the image reader's eye movement features predict the diagnostic interpretation of the case. The average classification performance of the feature-based classifier in predicting diagnostic interpretation

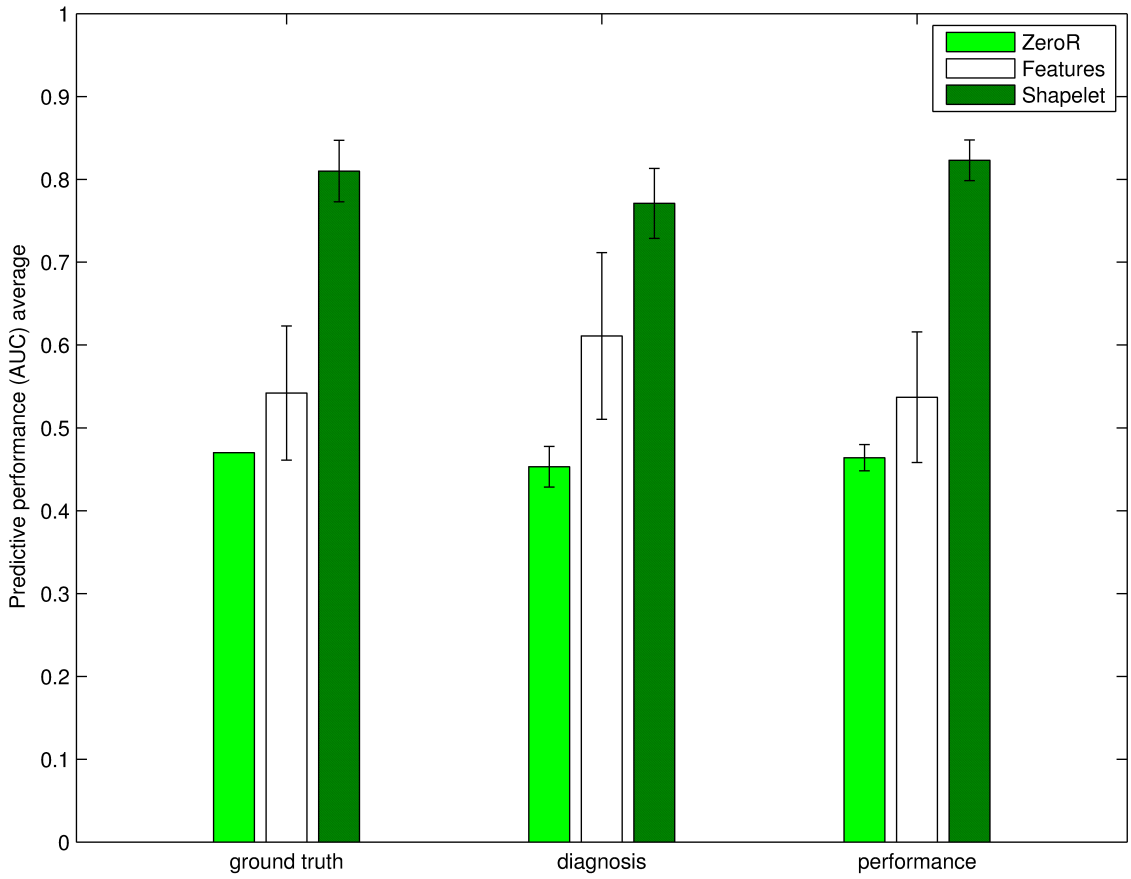


Figure 3.6: Aggregated performance results for predicting ground truth pathology, readers' diagnostic interpretation, and readers' performance.

(61%) was significantly higher ($p = 6e-4$) than results obtained from the baseline classifier (45%). The average classification performance the shapelet-based classifier (80%) was significantly higher ($p = 3e-4$) than the performance results using the feature-based classifier, and higher ($p = 9e-10$) than the performance results from the baseline classifier. The results from both classifiers, illustrated in Figure 3.6, indicate that pupillary responses can predict the image reader's impression or interpretation of the case being viewed with an accuracy significantly above random chance.

3.3.4.3 Predicting Image Reader's Diagnostic Performance

The purpose of this test was to measure the accuracy of predicting diagnostic performance using eye movement features. This test measures the ability of the classifier to predict whether or not the image reader will make the right diagnosis or not. The average accuracy obtained from the feature-based classifier in predicting image readers' performance (54%) was significantly higher ($p = .02$) than results obtained from the baseline classifier (46%). The results from the shapelet-based classifier (88%) was significantly higher ($p = 5e-7$) than the performance results from the feature-based classifier, and higher ($p = 5e-11$) than the performance results from the baseline classifier. The results from both classifiers, illustrated in Figure 3.6, show that pupillary responses can be used as a predictor for image reader's diagnostic performance.

3.3.5 Discussions

Prior research findings have shown that pupillary changes capture information about what is perceived, cognitive/mental processing, and cognitive/mental load. Deducing from these prior observations, we tested the efficacy of pupillary features as predictors of task performance. Our initial analysis (see Figure 3.5) showed sig-

nificant effects from eye movement features. These effects did not translate to high predictive performance on average using aggregate features, which are typical in machine learning algorithms. However, our results indicate that applying time series shapelet analysis on a single feature (change in pupil dilation), results in a significant improvement in predictive accuracy. Our findings in this study indicate:

1. The characteristics of a mammographic case (pathology and breast parenchyma density) are independent factors in predicting eye movement and pupillary changes during mammographic screening.

2. The experience level of the image reader is a significant factor in predicting eye movement and pupillary changes during mammographic screening.

3. The performance of aggregate features from eye-movement and pupillary changes marginally outperforms random chance at predicting case characteristics and diagnostic performance.

4. Measures such as time series shapelets, which capture temporal changes in eye movement and pupillary changes, perform significantly better than random chance at predicting case characteristics and diagnostic performance.

4. BIOMETRIC IDENTIFICATION OF IMAGE READER AND EXPERTISE IN RADIOLOGY

In this study, we present two methods for developing biometrics for identification of individual image readers and their corresponding experience levels from pupillary activity recorded while the latter performed a series of goal oriented tasks. This study focuses on the task of breast cancer detection in screening mammography clinically equivalent experimental conditions. Using a head-mounted eye-tracker, pupil and eye-movement data, and diagnostic interpretations for 100 mammographic cases of varying pathology were recorded for ten image readers with experience levels ranging from board certified radiologist to first year Radiology resident.

Saccadic eye-movements and pupillometric data were extracted from sensor recordings. We developed two methods to analyze these data for the purpose of extracting biometric identifiers and identifiers of experience level. The first method applied time-series shapelet analysis to build a dictionary of highly discriminative shapelets from pupillary change data. The second method applied sketch recognition algorithms to extract a set of geometric-based and gesture based features from saccadic eye-movement data. Each method was assessed independently by constructing predictive models for identifying the individual image reader and his/her experience level. We present the results from each method and analyze observed differences between both results.

Saccadic eye-movements, encoded using sketch gesture recognition features, are strong predictors of both individual identity and their corresponding experience level. Pupillary changes, encoded using time-series shapelets, bore out similar results as an equally effective predictor of individual identity and experience level. In addition,

pupillary changes proved useful in predicting the characteristics of a mammographic case, including case pathology and breast parenchyma density, image reader experience level. However, further analysis is needed to form conclusions on the effectiveness of a predictive model for mammography case characteristics and diagnostic performance using saccadic features from sketch recognition.

4.1 Introduction

Survival of breast cancer disease is largely dependent on early detection through the mammographic screening process. Studies show that through early detection, while the disease is localized, patients have a 98.5% relative survival rate versus 25% when the cancer is metastasized, a point at which the disease becomes incurable [247].

Early detection of breast cancer is made possible through mammographic cancer screening and diagnosis. Mammography is specialized medical imaging that uses low-dose X-ray to capture images of the interior tissues of the breast. While a mammogram can be used as a screening tool to detect early breast cancer in asymptomatic women, it can also be used to detect and diagnose breast disease in women experiencing symptoms such as a lump, pain, skin dimpling or nipple discharge. The former usage, screening mammography, is essential to early detection of breast cancers because it detects changes in the breast up to two years before any physically detectable symptoms appear. In contrast, diagnostic mammography is used to evaluate a patient with abnormal clinical findings, such as a breast lump or nipple discharge, detected through physical examination performed by the individual or a medical professional. Diagnostic mammography may also be performed after an abnormal screening mammogram, to evaluate the area of concern on the screening exam.

Both types of mammographic examination are performed by trained Radiologists. Radiologists are specialist physicians who utilize an array of advanced medical imag-

ing techniques to diagnose and, in some cases, treat patients with different types of diseases or injuries. Acquisition of expertise in radiology requires specialized training, consisting of 5 - 7 years of Radiology residency and fellowship, experience and a natural intuition for the task. Expert radiologists exhibit notably outstanding characteristics, such as increased speed and higher overall accuracy with which he/she makes decisions on the pathology of an image, which differentiate them from non-experts. However, the type and quantity of training and the experience necessary for expertise has been the subject of research in artificial intelligence [45].

Although the exact relationship between experience and expertise remains unclear, one approach to establishing a quantitative relationship between the number of years of experience and the attainment of expertise in mammography, is through identifying differences in visual search behavior between experts and non-expert image readers [193].

In a study of six image readers (three with 8 - 15 years of experience reading mammograms as board certified radiologists, and three with 3 - 4 years of experience reading mammograms as Radiology residents) Krupinski [139] compared cumulative cluster dwell times on 20 mammographic cases between both experience groups. A comparison of the median values for experienced and inexperienced image readers revealed that experienced readers tend to have shorter dwell times. Their findings suggest that temporal measures of visual search behavior may be important factors in differentiating experience level of image readers.

Kundel and LaFolette [150] evaluated the eye movements of 24 subjects, which included laymen, medical students, and experienced radiologists while viewing normal and abnormal chest radiographs. They reported an evolution of observers' scanpaths for chest radiographs from a localized central patterns of first-year medical students to circumferential patterns of the experienced radiologist. They noted that, in addi-

tion to the distinct nature of experienced radiologists' scanning patterns, experienced radiologists also moved their eyes to the target faster, and were more accurate in interpreting what they saw. Kundel and LaFolette's findings suggest that geometric properties of the scanning patterns formed during visual search may be important factors in differentiating between experienced and inexperienced image readers.

To investigate human factors associated with the proficiency of diagnostic pathology, Krupinski et al. [143] conducted a study examining the eye movements of nine image readers, belonging to one of three experience groups (medical students, Pathology residents, and pathologists). In this study, each of the nine slide readers took part in a single 45 minute virtual slide reading session, during which they examined 20 breast core biopsy slides while their eye movements were recorded. They found that experienced pathologists had the longest saccade length on average (measured in seconds) compared with residents, who in turn had longer saccade lengths on average compared with medical students.

In addition, they found that the average saccade velocity (measured as length per second) for experienced pathologists was significantly lower than residents, who's average saccade velocity was higher than the average velocity for medical students. They also reported that the decreasing trend in saccade velocity with years of experience was consistent within the experienced pathology group, with the most experienced pathologist having a significantly lower average saccade velocity than the less experienced pathologists. Krupinski et al's findings suggest that distance and velocity measures of eye movements during visual search may also be important factors in differentiating between experienced and inexperienced image readers.

4.2 Related Work

4.2.1 Sketch Gesture Recognition

Sketch is considered as a natural form of communication involving free form shapes, letters and numbers, which encode contextual meaning [204, 131]. Sketches can be considered as a special class of gestures. The fundament in sketch recognition involves encoding patterns contained within a sketch gesture in a manner that permits accurate interpretation and inference based on the intent of the author of the sketch gesture [104, 87]. Sketch recognition utilizes machine intelligence to capture and interpret intent of the author making the sketch gesture. The correct interpretation of gesture intent lets us to integrate sketch gestures into interfaces and systems, which in turn allow us to perform intelligent manipulation and computation on the recognized input.

There have been numerous algorithmic contributions in the field of sketch gesture recognition. The majority of sketch recognition algorithms fall into one of three broad categories [96]: geometry-based recognition algorithms [210], vision-based (appearance-based) recognition algorithms [127, 201], gesture-based (motion-based) recognition algorithms [234, 153], or their combination [55].

Geometric-based algorithms apply geometric relationships and constraints to describe primitive (basic) shapes [92, 94, 88, 89, 124, 102, 86, 262, 95]. Paulson and Hammond [210] developed a primitive sketch recognition and beautification system (*Paleosketch*). *Paleosketch* was designed to recognize sketches based on a bottom up approach of identifying low-level primitive shapes as components, which combine to form a recognizable high-level shape. In addition, the *Paleosketch* system returns a beautified version of the recognized shape. To achieve an accurate primitive sketch recognition in *Paleosketch*, Paulson and Hammond developed and tested two novel

features in the pre-recognition stage: the normalized distance between direction extremes (*NDDE*) and the direction change ratio (*DCR*) [100]. *NDDE* computes the the difference between the point of highest direction value and the lowest value normalized by stroke lengt. This metric differentiates curved shapes (high NDDE values) from poly-lines, which have lower NDDE values. *DCR* computes the ratio of maximum change in direction to the average change. This metric is useful for differentiating between poly-lines, which have a higher *DCR* and curves, which have a much lower value in comparison.

Appearance-based recognition algorithms rely on the appearance of a sketched shape; ignoring timing and ordering constraints of data points [226]. These algorithms rely on recognition techniques, such as template-matching, on the snapshot of a sketched shape (usually in the form of a bitmap) to distinguish between shapes. Kara and Stahovich [127] used a multi-layer recognition scheme to develop a trainable, hand-drawn symbol recognition system capable of recognizing hand-drawn symbols. The symbol recognition system uses a scaled bitmap image representation for sketched input. The bitmap image is then converted into polar coordinates to achieve rotational invariance. Unseen examples are classified based on the best aggregated similarity score returned by an ensemble of four classifiers (Hausdorff distance [235], modified Hausdorff distance [63], Tanimoto coefficient [73], and Yule coefficient[271]. They reported an accuracy of 95.7% and 94.7% on user-dependent and user-independent tests respectively.

Gesture-based (motion-based) recognition algorithms rely primarily on the path of motion of a stroke. These types of algorithms were initially conceptualized for the identification of a small set of application specific gesture commands [234, 153, 305, 246, 46, 155]. Gesture-based algorithms are able to characterize shapes based on how individual strokes are drawn (the path of each stroke) and not necessarily the shape of

the stroke, even though the latter can be correlated. In [234], Rubine presented a pen input gesture-based recognition system (GRANDMA), which enabled recognition of single stroke gestures through simple trainable linear classifiers. In this work, Rubine proposed and evaluated 13 features on classifying ten different gesture datasets, each comprised of 15 classes, and reported an average accuracy of 98%. In a followup work, Long et al. [153] proposed nine additional features to those developed by Rubine in [234]. Long et al. used multi-dimensional scaling models to identify correlated features in Rubine's feature set. Through their analysis, they identified an optimal feature set of 11 of Rubine's original 13 features, thereby eliminating two features. In addition, Long et al. proposed 11 new features, which combined with Rubine's features result in improved prediction of similarity between sketch gestures. Since gaze data primarily capture motion of the eyes, we expect gesture-based (motion-based) recognition algorithms will provide the highest utility for recognition purposes.

Sketch recognition algorithms were previously applied to solve challenging pattern recognition problems in other domains [263, 290, 99]. Dixon and Hammond in 2010 [59, 56, 60, 101] and Pramanik and Bhattacharjee in 2012 [222] use sketch recognition to identify faces in images from sketched drawings. In this work, they extracted geometric features from hand-drawn sketches and similar features from a set of images containing the same individual as a bases for comparison. Using k nearest neighbor (K-NN) classifier, they reported an average of 86% similarity with the top five matches using their method, which was much higher than averages from the two alternatives presented (eigenface: 43%, and sketch transform method: 80%).

Cig and Sezgin [51] developed an eye gaze movement interaction system, which can interpret eye movement patterns as auxiliary commands to augment pen-based gestures as a mode selection mechanism (drag, minimize, scroll etc.) during sketch interaction. The presented method tracks eye movements during pen gestures, com-

putes two features (gesture - gaze distance metric, and within-cluster variance of gaze positions), and uses these features to predict one of 5 tasks. Their results show that manipulation commands, which require auxiliary mode switching elements (such as multi-finger gestures), can be deduced with 88% accuracy using natural gaze behavior during pen interactions.

In [201], Ouyang and Davis presented a robust, multiple domain sketch recognition system, which uses vision based decomposition methods to classify hand-drawn symbols. Their system represents symbols as a set of feature images, in contrast to geometric or temporally ordered data points. These image features capture properties of the constituent strokes in a sketch symbol, such as orientation and the location of end points. Each sketch symbol is then classified using a deformable template matching algorithm based on image deformation model (IDM), which is robust to shifts and local noise distortions. Their system was evaluated and compared with 7 similar systems on datasets from three distinct domains (pen digits, HHReco[116], and circuit diagrams). They reported an average performance accuracy of 97.9%, with support vector machine with a radial basis function (SVM-RBF) kernel having the second best performance average (95.4%).

Several high level systems [97, 257, 105, 258, 98, 57, 256, 255, 254, 188, 276] use geometric algorithms to recognize shapes using the low level features that we use to recognize saccades. Valentine et al. presented *Mechanix*, an intelligent, interactive, on-line tutoring system, which allows engineering students to enter planar truss and free-body diagram solutions to homework problems [278, 72, 215, 275, 277, 9, 133, 106, 11, 10, 223, 12]. Mechanix uses recognition algorithms to segment [294, 289, 293, 291, 292] and recognize [206] primitive shapes (artifacts containing at most a single stroke). These low level primitive shapes are then systematically organized to form complex shapes using geometric-constraint-based recognition algorithms[90, 93, 103,

208]. These procedures give the Mechanix system the impressive ability to verify and compare student submissions against instructor provided template solutions and give students feedback in real-time.

This work is not the first time such features have been applied to human motions other than pen [205, 209, 179, 17, 259, 77, 216, 261, 260], but it is the first time they have been applied to recognize eye motions.

4.2.2 Eye Movement as a Biometric

Biometrics refer to authentication techniques that rely on easily verifiable physical characteristics of an individual. Biometric *identifiers* are categorized as measurable physiological and behavioral properties of the individual. Physiological characteristics are measures related to some property of the physical body, including but not limited fingerprint, footprint, palmprint, palm veins, face recognition, DNA, iris recognition, and retina. Behavioral characteristics are measures specific the behavior of a person (behaviometrics), including but not limited to typing rhythm, gait, hand-writing, and voice. Eye movements do not easily lend themselves to forgery, since they are largely dependent on brain activity and extra-ocular muscle characteristics, which are tied to the individual. This property makes eye movement an attractive option for biometric identification.

In [195], Noton and Stark, were first to observe that individuals tend to repeat certain scanpath trajectories during repeated viewings of a given pattern. In their experiments, Noton and Stark tested this theory, coined *scanpath theory*, and found that the general scanpath displayed by a subject during a first viewing of a pattern was repeated in the initial eye movements of approximately two-thirds (65%) of subsequent viewings. In addition, Noton and Stark observed that the scanpath produced by an individual for a given stimulus pattern is unique and varied for each

subject [195]. These findings were also supported by subsequent research inquiry focused primarily on reading related information processing [232, 241].

Eye-movements were first explored as a potential biometric identifier in [130]. In this work, Kasprowski and Ober used a combination of eye reaction time (the period of time between introduction of stimulus and eye reaction), and stabilization time (the time taken for the eye to fixate on a new location after stimulus), as features to build a predictive model. Using data from nine subjects, they tested four different classification models using 10-fold cross-validation (k-nearest neighbor, naive Bayes, C4.5 decision tree, and support vector machines) and reported the best average false acceptance rate of 1.48% achieved with KNN (k=3).

Subsequently, researchers have explored various eye-movement measures including: gaze trajectory [58, 75], gaze velocity [301], and pupil size [22] with reasonable success. Galdi et al. developed a gaze analysis based (GAS) soft-biometric predicated on user behavior during observation of a particular object, such as facial images [75]. The GAS system uses a fixed area of interest (AOI) based feature vector, which is derived using order-independent cumulative duration of fixation on the respective area of interest. The algorithm creates a profile of the observer by averaging corresponding AOI values over a set of images (16). The system is able to identify a test sample from an observer by computing the profile with the lowest Euclidean distance from the test sample. Galti et al. assessed the accuracy of their system on 88 test subjects and reported encouraging results in terms of receiver operating characteristic curves (ROC), equal error rate (EER), and cumulative match curve (CMC).

Yoon et al. explored the use of gaze as a biometric by examining the scanpath of 12 subjects viewing 50 images of patterns with varied spatial characteristics. They modeled gaze velocities using Hidden Markov Models to create unique profiles for each individual. Using a leave-one-out cross-validation evaluation, they reported an aver-

age performance accuracy in user identification ranging between 53% and 76% [301].

Holland and Oleg evaluated eye movement-based measures as features for biometric identification. They recorded the eye movements of 32 participants (26 male / 6 female) using a head mounted eye tracking device. Each participant's eye movements were recorded while performing a challenging reading task. From the recorded data, they extracted basic eye movement features and scanpath measures including: fixation count, fixation duration, saccade amplitude and velocity. Applying an information fusion method, they combined these features and reported a 27% error rate in identifying subjects [114].

4.3 Materials and Methods

4.3.1 Image Database

Table 4.1: Specifications of the 100 four-view screening mammograms used in the study.

Ground Truth	**Patient Age**	**Breast Density**	**Mass Subtlety**	**Total Abnormalities**	**No. of Cases**
Normal	Range: 36 68 (56.2 10.6)	Range: 1 4 (Median: 2)	N/A	N/A	25
Benign	Range: 34 82 (56.9 13.4)	Range: 1 3 (Median: 2)	Range: 3 5 (Median: 5)	Range: 1 3 (Median: 1)	25
Malignant	Range: 37 83 (64.3 12.4)	Range: 1 4 (Median: 2)	Range: 1 5 (Median: 5)	Range: 1 3 (Median: 1)	50

For the proposed study, we selected 100 screen-film mammograms from a corpus of mammographic images, digitized using a high resolution LUMISYS scanner (50m per pixel, 12 bit), sourced from the University of South Floridas Digital Database for Screening Mammography (DDSM) [110]. The DDSM database contains roughly 2,500 normal and abnormal cases. Each case within the DDSM database is accompanied by

associated patient information, craniocaudal (CC) and mediolateral oblique (MLO) view mammographic images of both the left and the right breasts. Abnormal cases are accompanied by duplicate images containing pixel level ground truth markings of abnormalities, and ground truth subtlety values using the BI-RADS lexicon [198] established via biopsy, additional imaging, or two-year follow-up.

Each of the 100 cases used in this experiment was manually selected to cover a broad range of mass margin and shape characteristics. Of these 100 cases selected, 50 included biopsy-proven malignant masses, 25 cases included biopsy-proven benign masses, and the remaining 25 cases were normal as determined during a two-year cancer-free follow-up patient evaluation. Masses with associated microcalcifications and malignant cases with benign lesions present were excluded during selection. Mass conspicuity was assessed according to the subtlety rating provided in the DDSM truth files. These ratings range from 1 (suggesting a subtle lesion) to 5 (suggesting an obvious lesion). A complete list of the DDSM cases used in this study is provided in the Appendix. Table 4.1 provides details on the selected cases, including information on the patients age and breast parenchymal density. The parenchymal density is also provided in the DDSM truth files, and ranges between 1 (fatty) to 4 (dense), according to the BI-RADS lexicon [198].

4.3.2 Experimental Procedure

Institutional review board approval was obtained prior to the study. We recruited ten readers of variable experience levels from an academic institution to conduct blind review of 100 four-view screening mammograms of varying pathology (see Table 4.2). Each reader was asked to report the location of any suspicious mass and provide a corresponding BI-RADS rating as typically done in clinical practice. Of the ten readers, three were experienced MQSA-certified radiologists each with at least

Table 4.2: Summary of characteristics of study participants.

Reader Type	**Experience Level**	**No. of Participants**
Radiologist	$>$ 10 yrs of practice	2
Radiologist	$<$ 10 yrs of practice	1
Advanced Resident	$>$ 2 mammo rotations	4
New Resident	$\leq$ 2 mammo rotations	3
Total		10

twelve years of dedicated mammographic experience, four radiology residents with at least two mammography rotations, and three radiology residents with at least one mammography rotation (see Table 4.2). Human subject recruitment and data collection was done according to a protocol approved by the Oak Ridge Site-Wide Internal Review Board. All participants signed an informed consent form.

A customized graphical user interface was developed in-house for study participants to view each mammographic case and record their findings. Two medical grade monitors were used (dual-head 5MP mammo-grade Totoku LCD monitors calibrated to the DICOM display standard). The four mammographic views (left, right, craniocaudal (CC), and mediolateral oblique (MLO) views) were initially displayed at low resolution (two views per monitor) to fit the screen. The GUI provided the functionality of zooming in/outs, panning, and magnifying glass for detailed reading of each mammographic view. During the reading sessions, each reader was outfitted with an H6 head-mounted eye-tracker, with a 60 Hz sampling rate, and eye-head integration from Applied Science Laboratories (ASL, Bedford, Massachusetts, USA). The eye-tracker recorded each readers eye position data to within 1 deg of accuracy. Prior to the study, each reader was carefully calibrated using the 9-point calibration protocol provided by ASL and trained using five training cases selected from the DDSM database.

Readers were instructed to take as much time as needed to view each case until they were satisfied with the viewing phase. Once the reader was prepared to give a diagnostic assessment of the case, the eye-tracking recording process was halted pending completion and reporting of case specific findings and they were ready to proceed with viewing the next case. The readers task was to mark and rate any finding suspicious for malignancy and any benign findings that they would normally report in clinical practice. Each mark was classified into a type of mammographic finding and rated for probability of malignancy on a BIRADS-based scale, which consists of five levels (2, 3, 4A, 4B, 4C, and 5) of increasing probability of malignancy. Cases with no markings were assigned a BIRADS rating of 1. On reading each case, the reader was instructed to proceed with the next case. After completing all the cases presented, data collection was discontinued and eye tracking equipment was removed. The experimenter subsequently debriefed and thanked the image reader for participating.

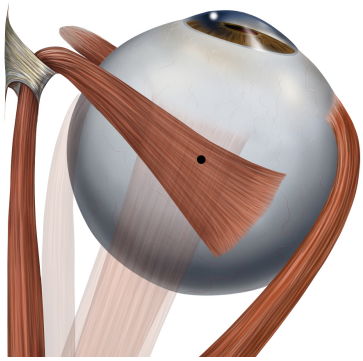
4.3.3 Eye Events

Eye events refer to voluntary or involuntary change in the configuration of the eyes such as movements, which help the subject to acquire, fixate or track visual stimuli. The movement of the human eye is controlled by pairs of muscles, who's combined and coordinated effect (depicted in Figure 4.1) is responsible for horizontal (yaw), vertical (pitch), and torsional (roll) eye movements, respectively; enabling them to control the three-dimensional orientation of the eye.

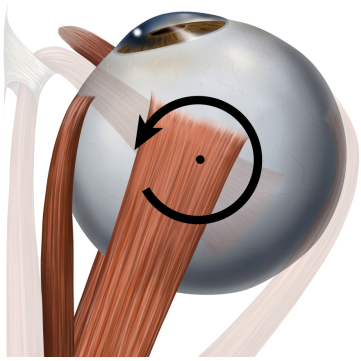
The three antagonistic pairs of muscles, which control eye movements are: the lateral and medial rectus muscles, the superior and inferior rectus muscles, and the superior and inferior oblique muscles. These muscles are responsible for movements of the eye along three different axes: horizontal adduction toward the nose or ab-

duction away from it; vertical elevation or depression; and intorsion or extorsion movements that bring the top of the eye toward or away from the nose respectively. The horizontal movement of the eyes are controlled entirely by the medial and lateral rectus muscles. The medial rectus muscle is responsible for horizontal adduction (towards the nose), the lateral rectus muscle for abduction (away from the nose). The vertical movements require the coordinated action of the superior and inferior rectus muscles, as well as the oblique muscles. The relative contribution of the rectus and oblique muscle groups depend on the horizontal position of the eye prior to the execution of this movement. In the primary position (eyes straight ahead), both of these groups contribute to vertical movements. The elevation of the eye results from actions of the superior rectus and inferior oblique muscle groups, while a depression of the eye results from actions of the inferior rectus and superior oblique muscle groups. However, when the eye is abducted, the rectus muscles become primary in the execution of vertical movement. Elevation results from the actions of the superior rectus muscle group, while a depression results from the actions of the inferior rectus muscle group. Conversely, when the eye is adducted, the oblique muscles become primary in the execution of vertical movement. Elevation results from the actions of the inferior oblique muscle group, while a depression results from the actions of the superior oblique muscle group. The oblique muscles are also primarily responsible for torsional movements [224].

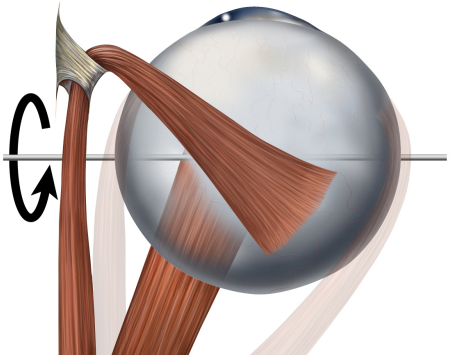
According to Donders law [273], orientation uniquely decides the direction of gaze, independent of how the eye was previously orientated. Large sections of the brain control these muscles to direct the gaze to the desired locations in space. Humans and other primates (including other vertebrates) primarily engage in seven types of voluntary and involuntary eye movement: fixation, saccade, glissade, smooth pursuit, microsaccade, tremor, and drift (see Table 4.3) [115]. We recorded gaze data for



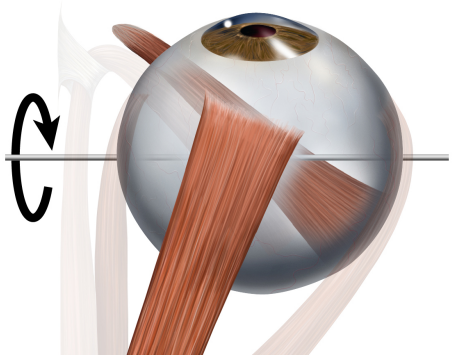
(a) Eye movement of lateral rectus muscle (From Lynch [157]).



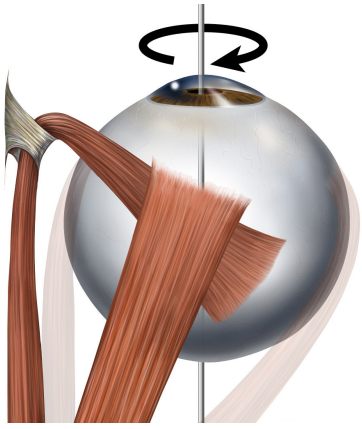
(b) Eye movement of medial rectus muscle (From Lynch [158]).



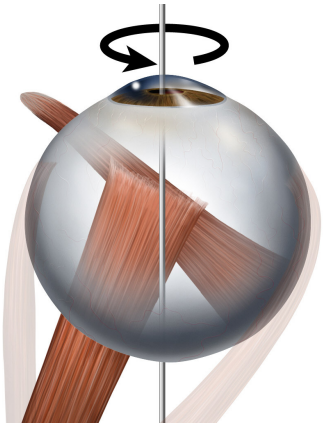
(c) Eye movement of inferior rectus muscle (From Lynch [159]).



(d) Eye movement of superior rectus muscle (From Lynch [160]).



(e) Eye movement of superior oblique muscle (From Lynch [161]).



(f) Eye movement of inferior oblique muscle (From Lynch [162]).

Figure 4.1: Superior view of muscles responsible for horizontal (yaw), vertical (pitch), and torsional (roll) eye movements.

each image reader from the four mammographic views (RMLO, LMLO, RCC, LCC) spread across two monitors (for each mammographic case). First, we processed raw gaze data to extract fixations. A *fixation* refers to a state where the eyes remain still (within a small radius) over a period of time, such as is the case when the eyes pause on a given word while reading. Fixating on a point or region is generally considered as a measure of attention to a given position or region of interest, even though this is not always the case.

Table 4.3: Summary of basic eye events

Description	**Duration (ms)**	**Amplitude**	**Velocity**
Fixation	$200 - 300$	N/A	N/A
Saccade	$30 - 80$	$4 - 20^{\circ}$	$30 - 500^{\circ}/s$
Glissade	$10 - 40$	$0.5 - 2$	$20 - 140^{\circ}/s$
Smooth Pursuit	N/A	N/A	$10 - 30^{\circ}/s$
Microsaccade	$10 - 30$	$10 - 40'$	$15 - 50^{\circ}/s$
Tremor	N/A	$< 1'$	$20'/s$ (peak)
Drift	$200 - 1000$	$1 - 60'$	$6 - 25'/s$

Although there are no universally excepted methods for computing fixations, there are parameters based on ocular physiology, which permit a reasonable criteria for detecting and extracting fixations from gaze data. A typical algorithm to determine a fixation event uses the mean X and Y eye position coordinates measured over a minimum period of time during which the eye does not move more than some maximum amount. This algorithm requires that a point-of-gaze must continuously remain within a small area (approximately within 1° visual angle in our algorithm) for some minimum amount of time (approximately 100ms for our algorithm). From this measure we computed the following features: rate of fixation (F_R), and the average duration of fixations (F_D) on a per case basis.

The eye is not completely still during a fixation, but exhibits three distinct types of micro-movements: *tremors*, *microsaccades*, and *drifts* [169]. A tremor is a small movement of approximately 90 Hz. The exact role of tremors is still a subject of active research; it is generally believed to be imprecise muscle control. Drifts are slow movements that shift the eye away from the centre of fixation, while the counter movement, a microsaccades, serves to quickly return the eye back to the center of fixation. However, these smaller, faster movements were not computed or utilized in this study.

The rapid motion of the eye from one fixation to another, from word to word while reading, for instance, is called a *saccade*. Saccades are considered the fastest movement the body can produce; typically taking 3080 ms to complete. It is a generally held view that human beings are perceptively blind during most of a saccadic event; a phenomena illustrated in Figure 1.4.

An important characteristic of saccades is that they rarely take the shortest path between two points, but instead undergo one of several *shapes* and *curvatures*. Since a saccade is described in terms of the gaze data between detected fixations, we computed saccadic events as gaze points connecting the completion of one fixation to the beginning of the next fixation. From the saccadic measure, we computed the following feature: saccadic amplitude (S_A). The *scanpath* is described as the eye-movement pattern that describes the route of occulomotor events through space within a defined timespan (such as the duration of a mammographic reading). From this measure, we compute the following feature: length of scanpath (SP_L).

The saccadic movement is not mechanically precise, that is they do not stop directly at the intended target, but instead the eye wobbles before coming to a stop. This post-saccadic movement is referred to as glissadic movements or *glissade*. The movement of the eye is characteristically different in the case of following or tracking

a moving object such as a bird flying across the sky. This type of eye movement is usually slower and referred to as *smooth pursuit*. The difference between the saccadic and the smooth pursuit movements is that the latter is driven by and requires a moving target, while the former can be made independent of any visual stimulus. This type of movement isn't evidenced in our experiment and was neither computed nor utilized.

4.3.4 Encoding Saccadic Movements

Once fixation and saccadic movement data were collected, we applied sketch recognition feature extraction algorithms to characterize the *shape* and *curvatures* of individual saccadic movements. Since the gaze scanpath is an aggregate shape consisting of individual saccadic movements, aggregating features extracted from saccadic movements will, in principle, provide an accurate characterization of the scanpath. Saccadic movements between displays (jumping from one display to another), thus between mammographic image views, were excluded from our computation. In addition, we computed similar features on the fixation scanpath. Features from the fixation scanpath were first computed independently for each mammographic view, then subsequently aggregated.

Gesture-based algorithms [151] are able to characterize shapes based on how individual strokes are drawn (the path of each stroke) and not necessarily the shape of the stroke, even though the latter can be correlated. Gestures can be user specific and even be used to differentiate users [68]. In [234], Rubine developed one of the earliest and arguably most recognized gesture recognition methods used for sketches. Using a set of stroke level features, Rubine trained a linear classifier for recognition of single stroke sketch gestures, and reported a high accuracy in differentiating between a small number of gestures (shapes) drawn as specified. Rubine's experiments reported

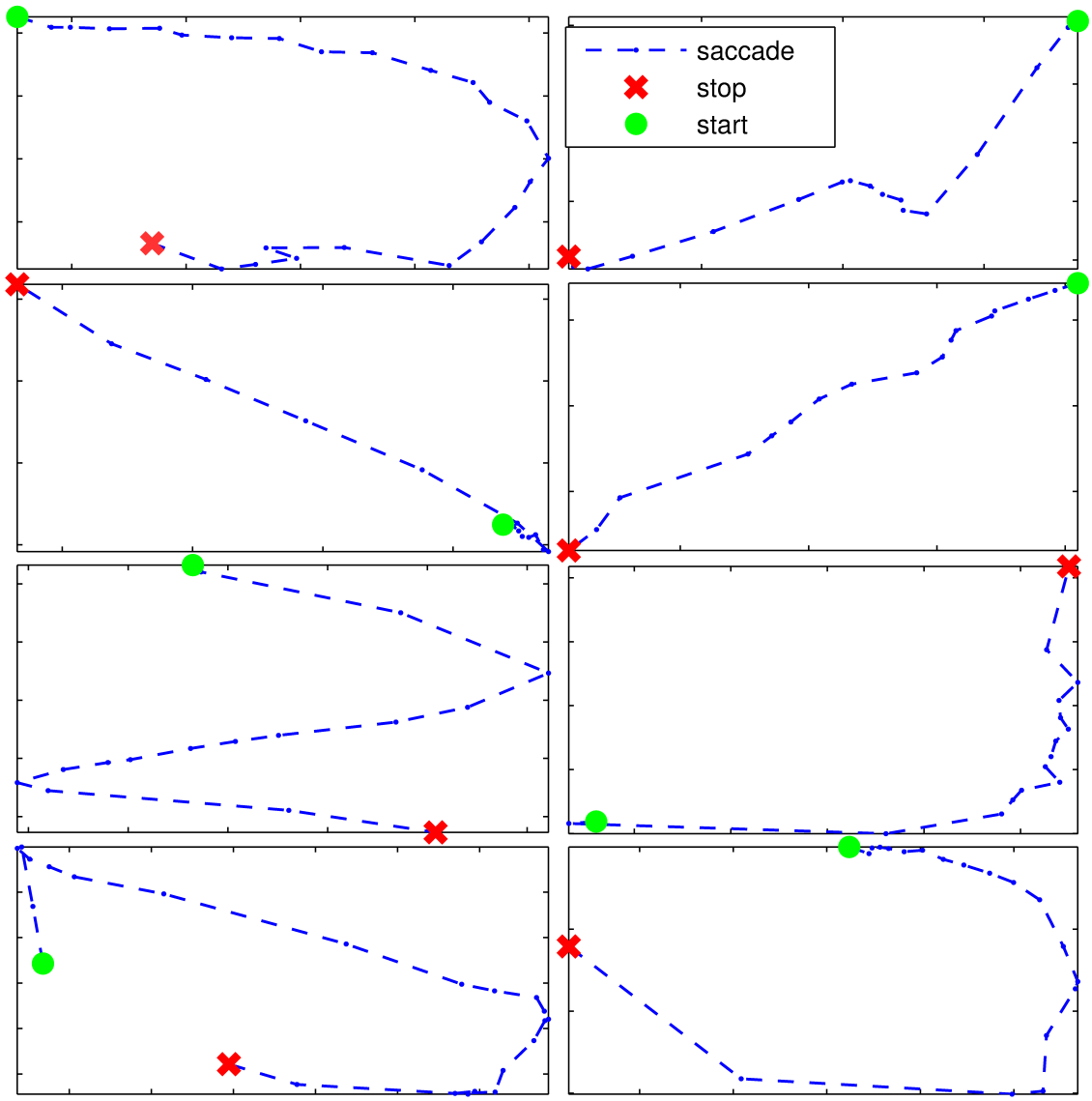


Figure 4.2: Sample saccadic movements recorded during a mammographic reading.

a total of 13 features as being important in the recognition of sketch gestures.

We selected 31 gesture-based and vision-based features, similarly to those examined in [211], which were previously demonstrated as being efficiently computable in real-time given a large input size, robust to noise, capable of encoding semantically meaningful information about a shape, and provide sufficient discriminative information to differentiate between shapes.

4.3.5 Rubine's Gesture Recognition Features

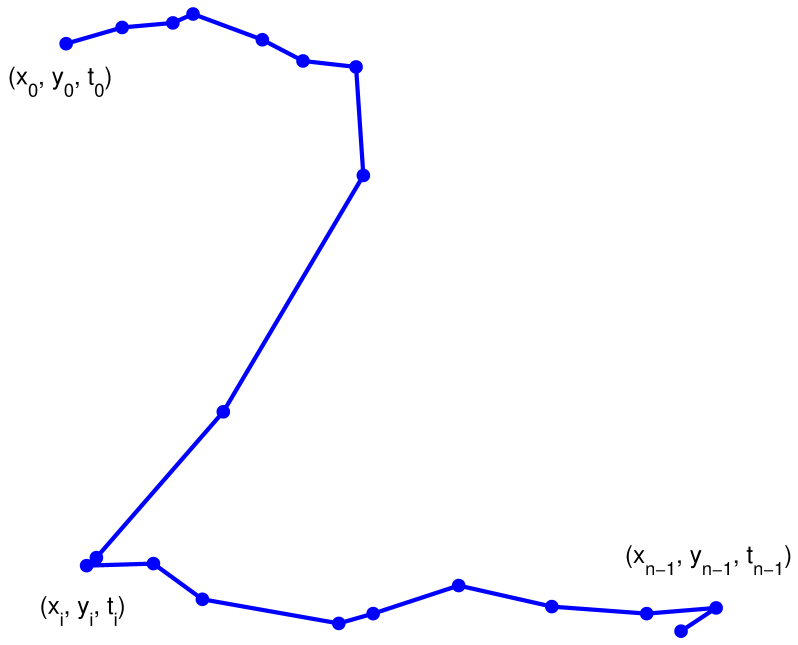


Figure 4.3: A sample saccade from a mammographic case reading. The x, y, and time values were sampled at 60Hz from a head-mounted eye tracking device.

Gesture-based algorithms are able to characterize shapes based on how individual strokes are drawn (the path of each stroke) and not necessarily the shape of the stroke,

even though the latter can be correlated. In [234], Rubine developed one of the earliest and arguably most recognized gesture recognition methods used for sketches. Using a set of stroke level features, Rubine trained a linear classifier for recognition of single stroke sketch gestures, and reported a high accuracy in differentiating between a small number of gestures (shapes) drawn as specified. Rubine's experiments reported a total of 13 features as being important in the recognition of sketch gestures. We describe each of the 13 features presented by Rubine.

First, in order to describe each feature, we will define the following terms as taken from [87]:

n	the total number of data points in a saccade
p_0	the first data point in a saccade
p_i	the i^{th} data point in a saccade
p_{n-1}	the last data point in a saccade
(x_0, y_0, t_0)	the x, y, and $time$ value for the first data point in a saccade
(x_{n1}, y_{n1}, t_{n1})	the x, y, time value for the last data point in a saccade
(x_i, y_i, t_i)	the x, y, time value for the i^{th} data point in a saccade
x_{min}	the minimum x value of the saccade (identical to the minimum x value of the bounding box)
x_{max}	the maximum x value of the saccade (identical to the maximum x value of the bounding box)
y_{min}	the minimum y value of the saccade (identical to the minimum y value of the bounding box)

y_{max}	the maximum y value of the saccade (identical to the maximum y value of the bounding box)
α	the starting angle of the saccade
β	the angle between the first (p_0) and last (p_{n-1}) point
θ_i	the angle of the line between the i^{th} and the $(i+k)^{th}$ point, for some constant offset k
d	the length of the bounding box encapsulating the entire saccade

4.3.5.1 Features 1 & 2 (f_1 & f_2)

Encode the starting angle of the saccade using a horizontal line as reference. Feature f_1 represents Rubine's first feature, which captures the cosine of the starting angle of the saccade, while feature f_2 represents Rubine's second feature, which captures the sine of the starting angle of the saccade.

$$f_1 = \cos(\alpha) = \frac{(x_2 - x_0)}{\sqrt{[(y_2 - y_0)^2 + (x_2 - x_0)^2]}} \tag{4.1}$$

$$f_2 = \sin(\alpha) = \frac{(y_2 - y_0)}{\sqrt{[(y_2 - y_0)^2 + (x_2 - x_0)^2]}} \tag{4.2}$$

4.3.5.2 Features 3 & 4 (f_3 & f_4)

Encode the bounding box of the saccade. Feature f_3 represents Rubine's third feature, which captures the length of the diagonal of the bounding box of the saccade.

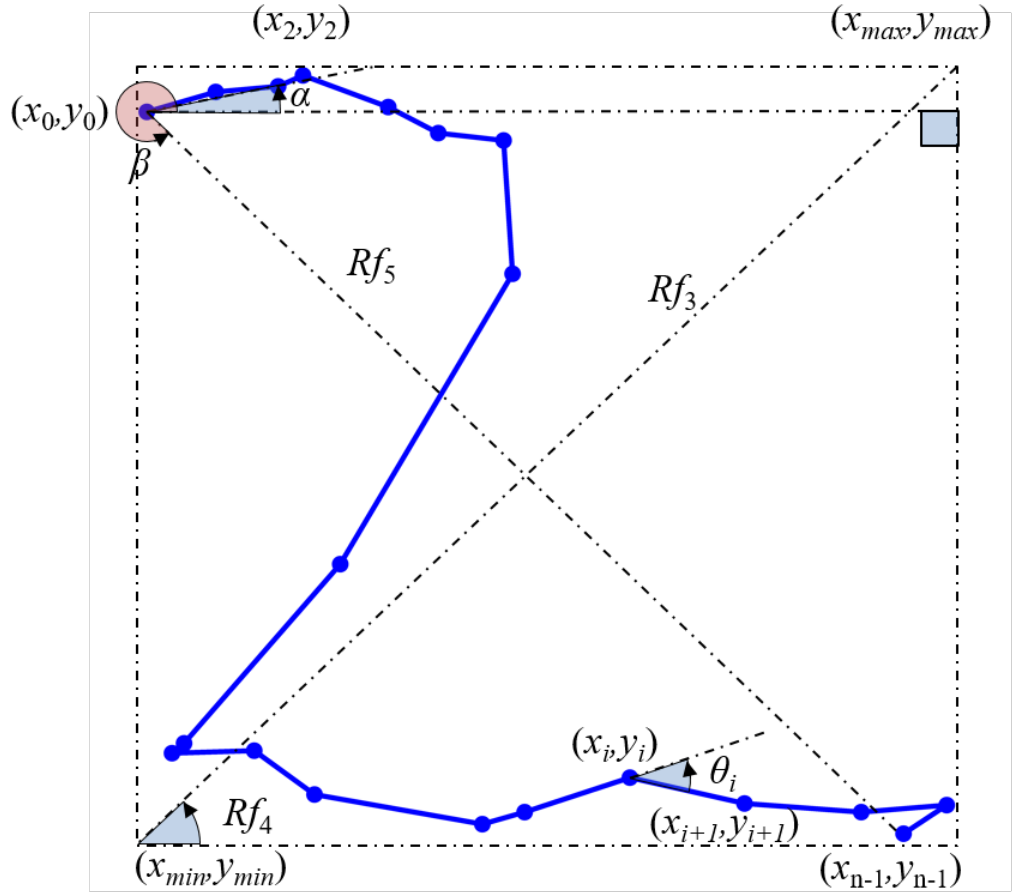


Figure 4.4: Rubine's features capture a multitude of properties associated with the shape of a saccade. Here, we visualize several of them

Feature f_4 represents Rubine's forth feature, which captures the angle of the diagonal of the bounding box. This feature measures the shape of the bounding box of a saccade.

$$f_3 = \sqrt{[(y_{max} - y_{min})^2 + (x_{max} - x_{min})^2]} \tag{4.3}$$

$$f_4 = \arctan \frac{\left[(y_{max} - y_{min})\right.}{\left.(x_{max} - x_{min})\right]} \tag{4.4}$$

4.3.5.3 Feature 5 (f_5)

Captures the distance between the start and the end point of the stroke. Feature f_5 represents Rubine's fifth feature, which is useful for differentiating closed saccades, such as circles from non-closed saccades, such as lines.

$$f_5 = \sqrt{(x_{n-1} - x_0)^2 + (y_{n-1} - y_0)^2} \tag{4.5}$$

4.3.5.4 Features 6 & 7 (f_6 & f_7)

Encode the angle between the horizontal line and the line formed by the first and the last point of the saccade (see β in Figure 4.4). Feature f_6 represents Rubine's sixth feature, which captures the cosine, while Feature f_7 represents Rubine's seventh feature, which captures the sine of the angle between the horizontal line and the line formed by the first and the last point of the saccade.

$$f_6 \;=\; \cos(\beta) \;=\; \frac{(x_{n-1} - x_0)}{f_5} \tag{4.6}$$

$$f_7 \;=\; \sin(\beta) \;=\; \frac{(y_{n-1} - x_0)}{f_5} \tag{4.7}$$

Feature 8 (f_8). Measures the total length of the path of a saccade. Feature f_8 represents Rubine's eighth feature, which is commonly described as the saccade length and is calculated by computing the euclidean distance between consecutive points, and summing all of these distances together.

$$f_8 \;=\; \sum_{i=1}^{n-1} \sqrt{\Delta x_i^2 + \Delta y_i^2} \tag{4.8}$$

where $\Delta x_i \;=\; x_i - x_{i-1}$ *and* $\Delta y_i \;=\; y_i - y_{i-1}$.

This feature helps to differentiate between saccades with similar bounding boxes, but where one saccade has significantly more movement.

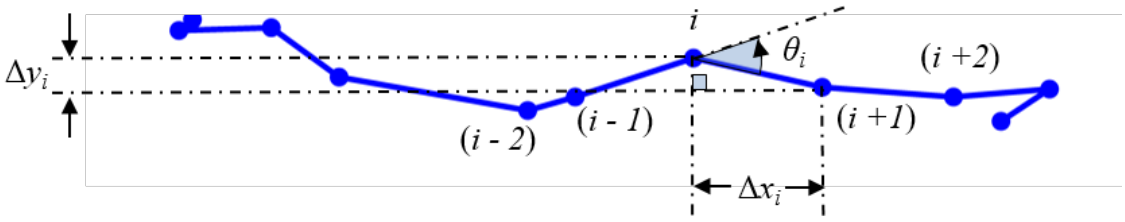


Figure 4.5: The θ_i value for a point p_i on a saccade.

4.3.5.5 Features 9, 10, & 11 (f_9, f_{10} & f_{11})

Encode the rotation of a saccade. Recall the definitions from above for the change in x and the change in y: $\Delta x_i = x_i - x_{i-1}$ *and* $\Delta y_i = y_i - y_{i-1}$. The angle between two points θ_i (see Figure 4.4 and Figure 4.5) is defined as:

$$\theta_i = \arctan\left(\frac{\Delta x_i \Delta y_{i-1} - \Delta x_{i-1}\Delta y_i}{\Delta x_i \Delta x_{i-1} + \Delta y_i \Delta y_{i-1}}\right) \tag{4.9}$$

Note that the *arctan* function ranges between $\pi/2$ and $\pi/2$ and is not continuous.

Feature f_9 represents Rubine's ninth feature, which computes the total rotational change in a saccade. This feature is computed by summing together the angles between each point on the saccade.

$$f_9 = \sum_{i=1}^{n-2} \theta_i \tag{4.10}$$

Feature f_{10} represents Rubine's tenth feature, which computes the total absolute rotational change in a saccade. This tenth feature is computed by summing together the absolute value of the angles between each point on the saccade. This feature encodes the degree to which the stroke moves around.

$$f_{10} = \sum_{i=1}^{n-2} |\theta_i| \tag{4.11}$$

Feature f_{11} represents Rubine's 11^{th} feature, which computes the smoothness of the stroke. This feature is computed by summing the square of the absolute values

of the angles between each point on the saccade. This feature encodes the degree of sharpness of turns during the saccadic movement. A saccade that has the appearance of a V for example, and an arc will provide similar values for features f_9 and f_{10}, but will provide different values for feature f_{11}. This differentiability correctly identifies saccadic shapes having sharp corners (such as squares) from saccadic shapes with softer corners (such as circles).

$$f_{11} = \sum_{i=1}^{n-2} |\theta_i|^2 \tag{4.12}$$

4.3.5.6 Features 12 & 13 (f_{12} & f_{13})

Encode the velocity of a saccadic movement. Each stroke point comes with a time stamp, where $\Delta t_i = t_i - t_{i-1}$.

$$f_{12} = \max_{i=1}^{n-1} \left[\frac{\Delta x_i^2 + \Delta y_i^2}{\Delta t_i^2} \right] \tag{4.13}$$

Feature f_{12} represents Rubine's twelfth feature, which computes the maximum speed reached during a saccadic movement.

Feature f_{13} represents Rubine's 13^{th} feature, which computes the total time duration of a saccade from start to finish.

$$f_{13} = t_{n-1} - t_0 \tag{4.14}$$

4.3.6 Long's Gesture Recognition Features

In [153], Long et al. developed a gesture recognition system similar to Rubine's. Long et al. used multi-dimensional scaling models to identify correlated features in Rubine's feature set and, through their analysis, identified an optimal feature set of 11 of Rubine's original 13 features, thereby eliminating two of Rubine's features in their final feature set. In addition, Long et al. proposed 11 new features, which combined with Rubine's features result in improved prediction of similarity between sketch gestures.

Feature 14 ***(f_{14})***. Calculates the aspect ratio of the angle of the diagonal of the bounding box. Feature f_{13} represents Long's 12^{th} feature, which is determined by calculating how much the angle of the diagonal of the bounding box deviates from 45°.

$$f_{14} = |45^\circ - \arctan\left[\frac{(y_{max} - y_{min})}{(x_{max} - x_{min})}\right]| \tag{4.15}$$

4.3.6.1 Feature 15 (f_{15})

Represents Long's 13^{th} feature, which calculates the curviness of a saccade in a somewhat similar fashion as Rubine's tenth feature (f_{10}), with the difference that it computes angles whose values are less than 19°.

$$f_{15} = \sum_{i=1}^{n-2} \begin{cases} |\theta_i|, & \text{if } \theta_i < 19^\circ \\ 0, & \text{otherwise} \end{cases} \tag{4.16}$$

4.3.6.2 Feature 16 (f_{16})

Represents Long's 14^{th} feature, which is a measurement of relative rotation, with the total rotation scaled by the saccade length. It is computed by dividing the total angle traversed by the total stroke length.

4.3.6.3 Feature 17 (f_{17})

Represents Long's 15^{th}, which is described as a density metric, comparing the saccade length (f_8) with the distance between the endpoints (f_5).

$$f_{17} = \frac{\sum_{i=1}^{n-1}\sqrt{\Delta x_i^2 + \Delta y_i^2}}{\sqrt{[(x_{n-1} - x_0)^2 + (y_{n-1} - y_0)^2]}} \tag{4.17}$$

4.3.6.4 Feature 18 (f_{18})

Represents Long's 16^{th} feature, which is described as another density metric, measuring the ratio of the saccade length (f_8) to the length of the diagonal of the bounding box (f_3).

$$f_{18} = \frac{\sum_{i=1}^{n-1}\sqrt{\Delta x_i^2 + \Delta y_i^2}}{\sqrt{[(y_{max} - y_{min})^2 + (x_{max} - x_{min})^2]}} \tag{4.18}$$

where $\Delta x_i = x_i - x_{i-1}$ *and* $\Delta y_i = y_i - y_{i-1}$.

4.3.6.5 Feature 19 (f_{19})

Represents Long's 17^{th} feature, which refers to a measure of non-subjective openness. Long considered (f_3) to be a subjective measure of openness of a saccade,

because it specifies the distance between the start and end points. By comparing the openness (f_3) of a shape to the size of the bounding box (f_5), the ratio provides a more relative measure of the distance between the two endpoints or the openness of a shape of the saccade.

$$f_{19} = \frac{\sqrt{[(x_{n-1} - x_0)^2 + (y_{n-1} - y_0)^2]}}{\sqrt{[(y_{max} - y_{min})^2 + (x_{max} - x_{min})^2]}} \tag{4.19}$$

4.3.6.6 Feature 20 (f_{20})

Represents Long's 18^{th} feature, which simply computes the area of the bounding box.

4.3.6.7 Feature 21 (f_{21})

Represents Long's 19^{th} feature, which computes the log of the area of the bounding box. Taking the logarithm helps to compress extreme values of area of the bonding box to a less extreme distance metric.

4.3.6.8 Feature 22 (f_{22})

Represents Long's 20^{th} feature, which computes the ratio of the total rotational change (f_9) to the rotational motion (f_{10}).

$$f_{22} = \frac{\sum_{i=1}^{n-2} \theta_i}{\sum_{i=1}^{n-2} |\theta_i|} \tag{4.20}$$

4.3.6.9 Feature 23 (f_{23})

Represents Long's 21^{st} feature, which computes log of the length of the saccade (f_8). As noted earlier, taking the log of of values results in a compression, making large values more similar to each other and small values more distinct.

$$f_{23} \;=\; \log[\sum_{i=1}^{n-1} \sqrt{\Delta x_i^2 + \Delta y_i^2}] \tag{4.21}$$

where $\Delta x_i \;=\; x_i - x_{i-1}$ *and* $\Delta y_i \;=\; y_i - y_{i-1}$.

4.3.6.10 Feature 24 (f_{24})

Represents Long's 22^{th} feature, which computes log of the aspect ratio of the bounding box (f_{13}).

$$f_{24} \;=\; \log[|45^\circ - \arctan\left[\frac{(y_{max} - y_{min})}{(x_{max} - x_{min})}\right]|] \tag{4.22}$$

4.3.7 Paulson and Hammond's Gesture Recognition Features

Geometric-based algorithms apply geometric relationships and constraints to describe primitive (basic) shapes. In [210, 212], Paulson and Hammond developed a primitive sketch recognition and beautification system (*Paleosketch*). *Paleosketch* was designed to recognize sketches based on a bottom up approach of identifying low-level primitive shapes as components, which combine to form a recognizable high-level shape. In addition, the *Paleosketch* system returns a beautified version of the recognized shape. To achieve an accurate primitive sketch recognition in *Paleosketch*, Paulson and Hammond developed and tested two novel features in the

pre-recognition stage: the normalized distance between direction extremes ($NDDE$) and the direction change ratio (DCR) [207]. In addition to Paulson and Hammond's proposed features, we included the average direction of the saccade (f_25), and the normalized curvature of the saccade (f_26). These features have been shown to help detect children's ages as well as corners [135].

4.3.7.1 Paulson's 1^{st} Feature (f_{27})

Computes the Normalized distance between direction extremes (NDDE). NDDE cacluclates the difference between the point of highest direction value and the lowest value normalized by saccade length (f_8). This metric differentiates curved shapes (high NDDE values) from poly-lines, which have lower NDDE values.

$$f_{27} = \frac{\max_{i=1}^{n-2} \frac{\Delta y_i}{\Delta x_i} - \min_{i=1}^{n-2} \frac{\Delta y_i}{\Delta x_i}}{\sum_{i=1}^{n-1} \sqrt{\Delta y_i^2 + \Delta x_i^2}} \tag{4.23}$$

where $\Delta x_i = x_i - x_{i-1}$ *and* $\Delta y_i = y_i - y_{i-1}$.

4.3.7.2 Paulson's 2^{nd} Feature (f_{28})

Computes the direction change ratioDCR computes the ratio of maximum change in direction to the average change. This metric is useful for differentiating between poly-lines, which have a higher DCR and curves, which have a much lower value in comparison.

$$f_{28} = \frac{\max_{i=1}^{n-2} \frac{\Delta y_i}{\Delta x_i}}{\sum_{i=1}^{n-2} \left[\frac{\Delta y_i}{\Delta x_i} \right] / n - 2} \tag{4.24}$$

where $\Delta x_i = x_i - x_{i-1}$ *and* $\Delta y_i = y_i - y_{i-1}$.

4.3.8 Alamudun and Hammond's Vision-Based Gesture Recognition Features

4.3.8.1 Retinal Activation

Retinal activation explains the tendency of the image readers' saccadic scanpath to follow a specific direction related to cortical maps. Cortical maps are collections (areas) of the brain identified as being responsible for processing a specific type of information. Neurons in the visual cortex are grouped together based on similar response properties that represent stimulus features such as edge orientation, direction of motion, and position in space. The intuition in the retinal activation feature is capturing individual behavioral adaptations resulting in a preferred overall direction of scanning. This value is computed as the average direction point to point movement in a saccade (f_{29}), found its cardinal orientation (f_{30}) and mapped this direction to one of 12 evenly spaced angles (30°, similar to five minute marks on a clock).

4.3.9 Time Series Shapelets

Machine learning classification algorithms have generated a significant amount of interest, and recent advances have provided efficient and accurate algorithms that solve difficult classification problems. However, much of the existing algorithms are designed for insulate numerical or symbolic values, where the relationship between numeric values (temporal, geometric properties etc) are not exploited.

This approach to solving classification problems works well for a restricted class of

problems, but generalize poorly. In the case of real-valued, ordered time series data, it becomes necessary to take into account such relationships as the temporal ordering and trends within the data, because they provide additional, and in most cases, critical discriminate information. These temporal information are not captured by aggregate (global) features, which are characteristic of machine learning classification algorithms.

Furthermore, while rigorous tweaks and optimizations to existing techniques may give satisfactory results in many cases, such as nearest neighbor algorithm, where features are individual data points and temporal characteristics are not preserved. A lot of improvement can be made by developing algorithms specific to the nature of the data. The resulting algorithms can not only provide more optimal and more accurate solutions, but in addition, the resulting output of these algorithms will be more interpretable since it encapsulates important temporal trends in the data.

These challenges point to the need for a different type of data primitive, which (a) capture temporal changes of observed data, (b) generate temporal attributes effective for establishing sufficient criteria for class membership and thus usable as a feature for classification, and (c) can be utilized during post-analyzed for model characteristics responsible for determining class membership.

One such primitive was recently introduced by Ye and Keogh [299]. In their work on data mining images in historical documents, Ye and Keogh developed a novel temporal primitive called a *time series shapelet*. They define shaplets as time series sub-sequences, which are maximally representative of a class. Intuitively, shapelets are features that capture temporal changes in data, and which can improve a model's ability to discern between two or more classes.

To illustrate the concept of time series shapelets, Ye and Keogh considered a two-class classification problem of identifying two commonly confused plants, *Uritica*

dioica (often called common nettle or stinging nettle) and *Verbena uriticifolia* (white vervain). Converting each sample into a one-dimensional representation,creates a data representation that can be used for classification using one of many existing techniques. Traditional techniques are shown to perform poorly on this type of problem [299] predominantly because the differences between the two classes are captured in the temporal changes in their respective data representation. However, a shapelet subsequence enables the comparison of temporal characteristics of both classes to successfully discriminate between the two. Using Ye and Keogh's methods [299], we use timeseries shapelet analysis to characterize temporal properties in image readers' pupil dilation during a mammographic screening [4].

4.4 Analysis and Results

4.4.1 Univariate Feature Analysis

We performed a univariate analysis to understand the underlying structure and characteristics of each feature, and ascertain its utility as part of a predictive model for biometric identification. First, we examined the distribution using a histogram to analyze the frequency distribution, and an aggregate plot showing the average and standard deviation for each feature. Figures 4.6 - 4.11 render a visualization of the 30 features extracted from our dataset. The histograms in Figures 4.6(a) - 4.11(a) show the range and distribution of values for each feature. This further illustrated in Figures 4.6(b) - 4.11(b), which show the averaged value and standard deviation of each feature value for each image reader. Notably, Figures 4.6(b) - 4.11(b) show some variation in the average value of each feature across the ten image readers.

Since the response (dependent) variable is nominal, we measured the value of each feature using a combination of model-based and gain ratio-based ranking. To compute the model-based rankings, we used the Random Forest classifier [37] algorithm

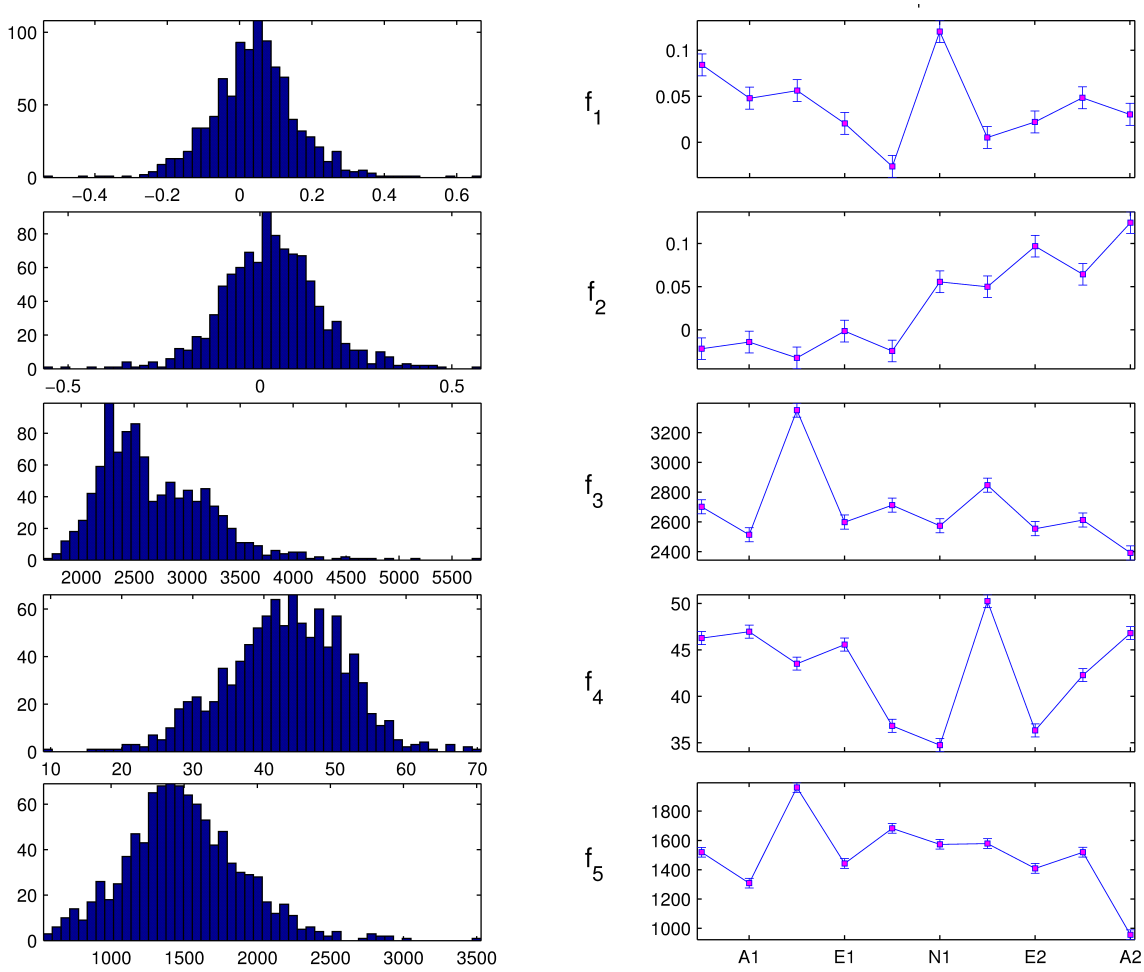


Figure 4.6: Illustrates the distribution of a subset of features in our dataset. (a) Histogram showing distribution of features across all image readers. (b) Average and standard deviation of features for each image reader.

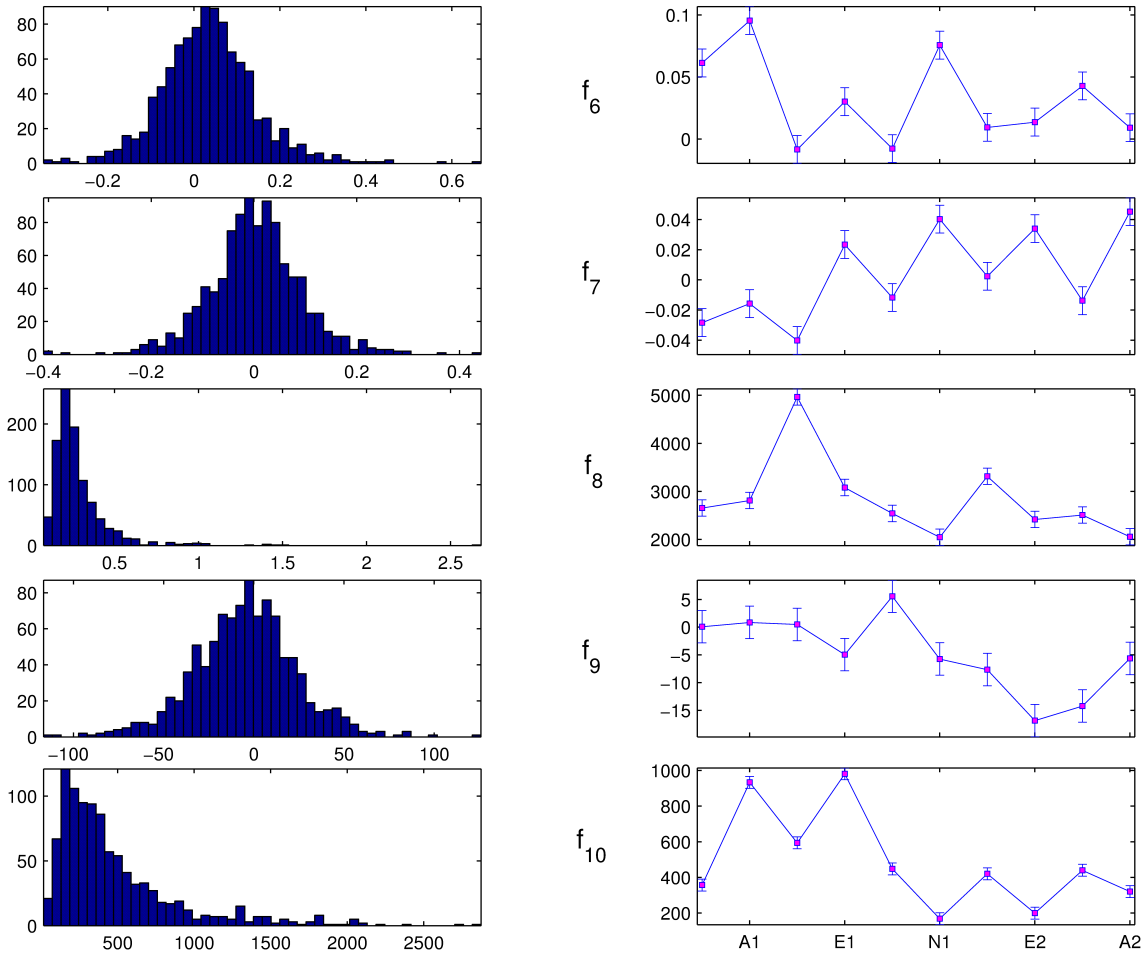


Figure 4.7: Illustrates the distribution of a subset of features in our dataset. (a) Histogram showing distribution of features across all image readers. (b) Average and standard deviation of features for each image reader.

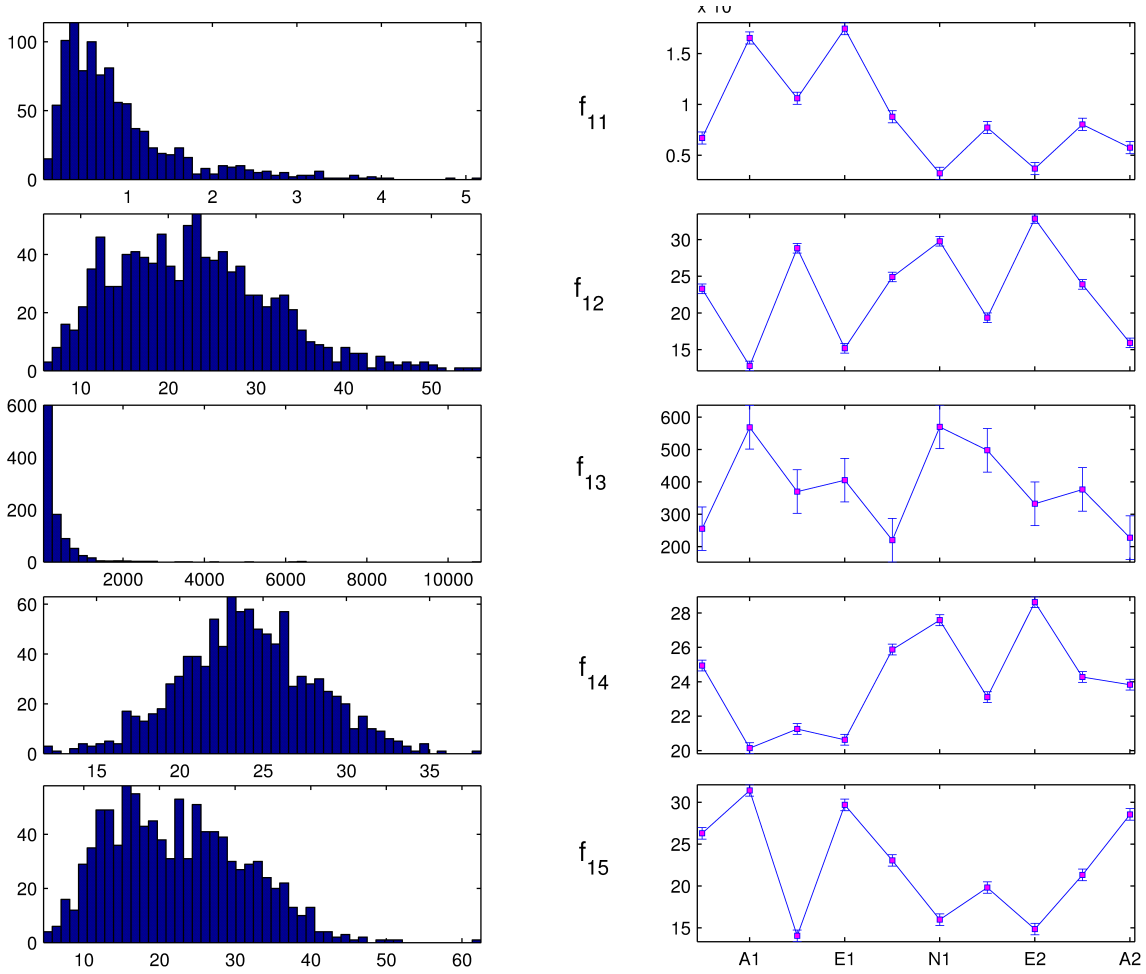


Figure 4.8: Illustrates the distribution of a subset of features in our dataset. (a) Histogram showing distribution of features across all image readers. (b) Average and standard deviation of features for each image reader.

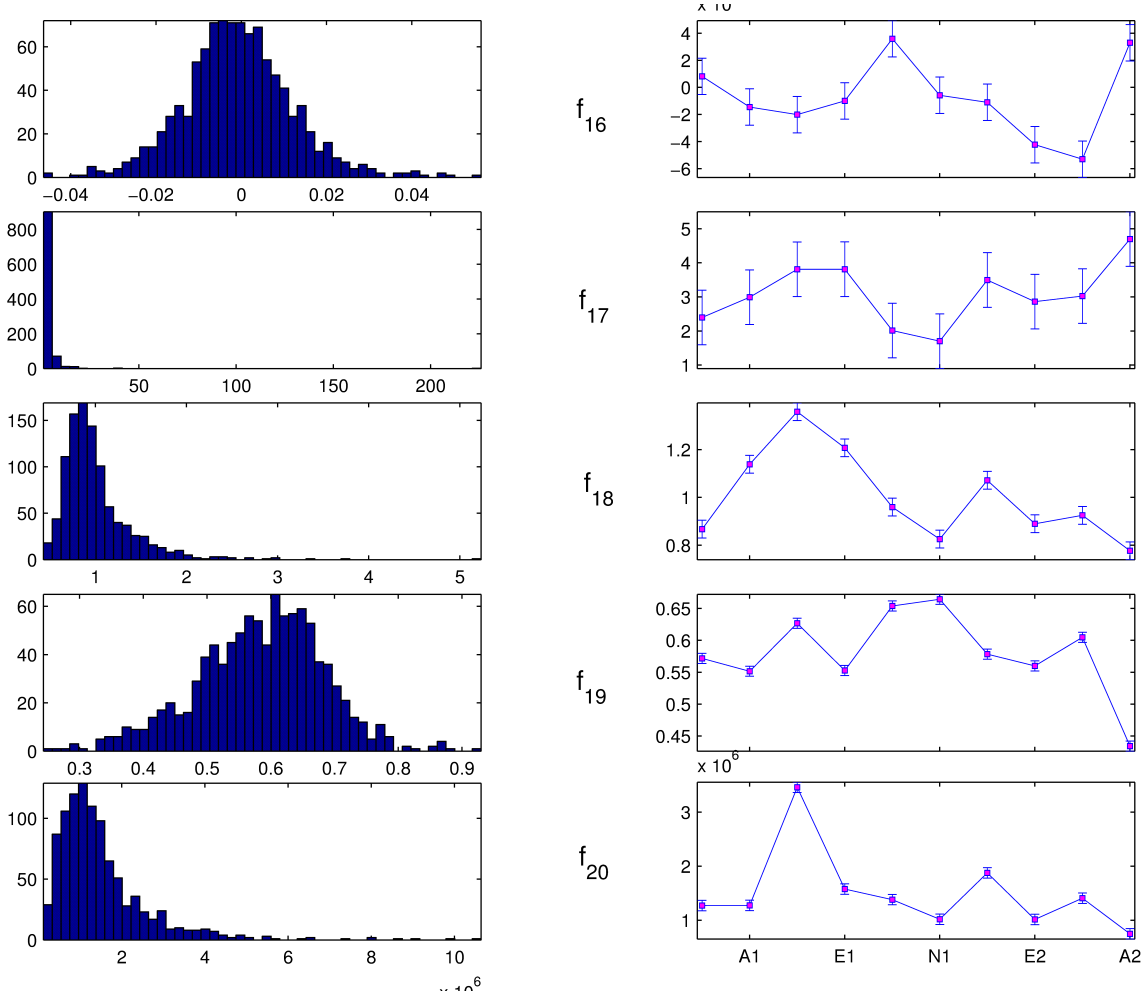


Figure 4.9: Illustrates the distribution of a subset of features in our dataset. (a) Histogram showing distribution of features across all image readers. (b) Average and standard deviation of features for each image reader.

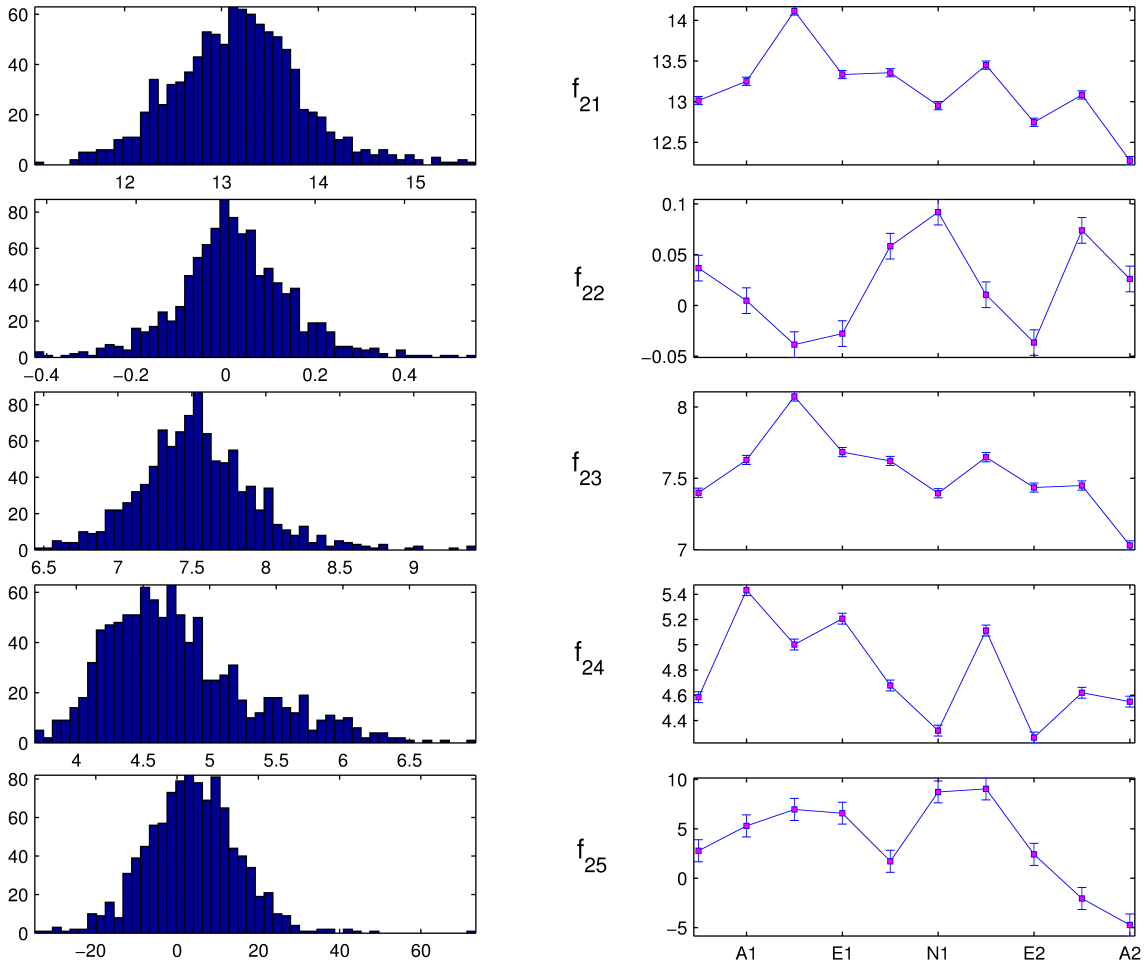


Figure 4.10: Illustrates the distribution of a subset of features in our dataset. (a) Histogram showing distribution of features across all image readers. (b) Average and standard deviation of features for each image reader.

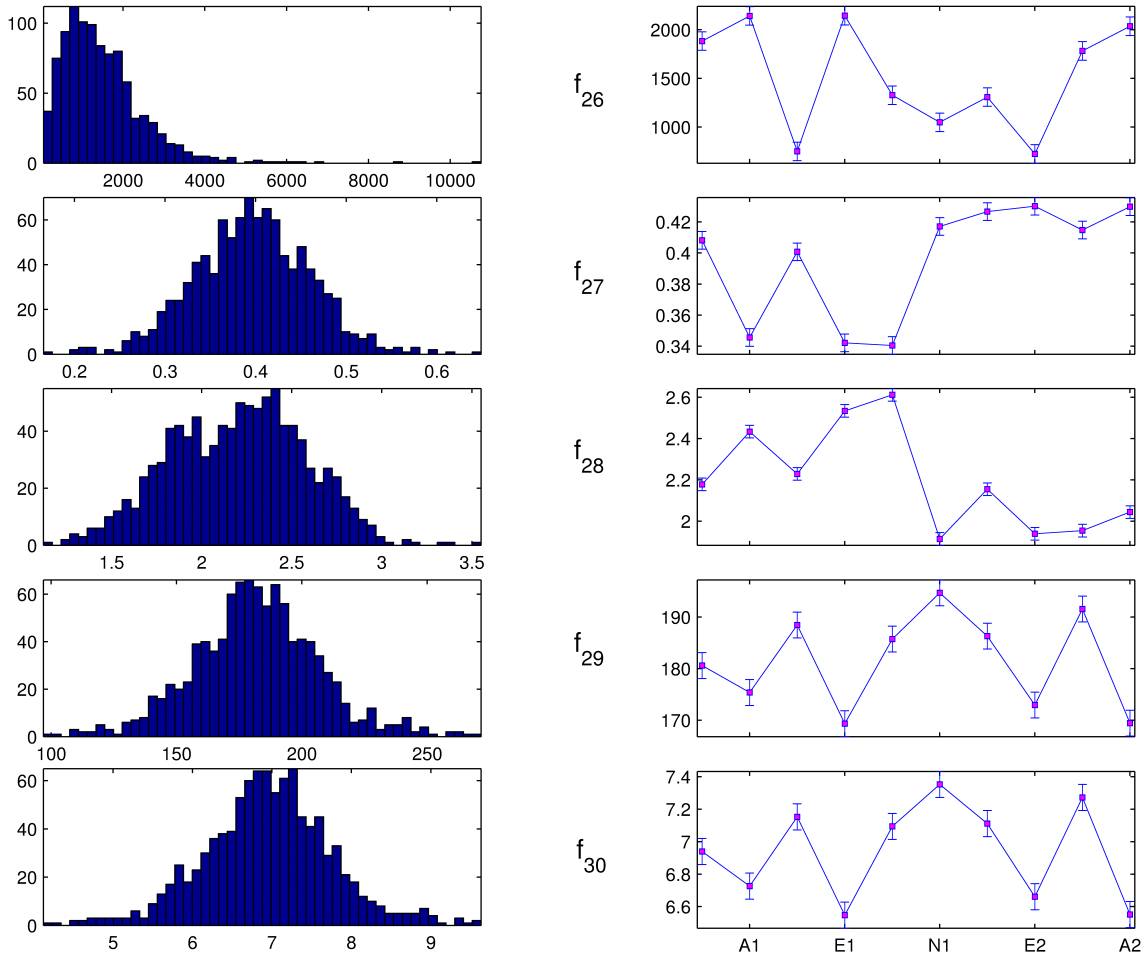


Figure 4.11: Illustrates the distribution of a subset of features in our dataset. (a) Histogram showing distribution of features across all image readers. (b) Average and standard deviation of features for each image reader.

to build 30 predictive models of image reader identity (one for each feature). Each predictive model was evaluated using a k-fold cross-validation scheme ($k = 10$). K-fold cross-validation involves partitioning the data into complementary subsets, performing analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). Multiple (k) rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds in order to reduce variability. The aggregated (mean) predictive value over all rounds serves as the final performance evaluation of the predictive model. Features were ranked by sorting according to their respective predictive accuracy in descending order.

Information gain is the expected reduction in entropy resulting from a partitioning of the dataset based on the values of a given feature. It quantifies the amount of information deducible about the response variable, based on the values of a given feature. However, information gain is not normalized and can therefore be biased in favor of large-valued features. The information gain ratio (IGR) resolves these limitations, reducing bias towards large-valued attributes, by incorporating the number and size of partitions into account when choosing an attribute. This process corrects the information gain by taking the intrinsic information of a split into account. We computed information gain for each feature with respect to the target class (image reader) and rank them in descending order.

In Table 4.4, we present the top ten features selected from both ranking algorithms. Next, we selected those features that appear in the top ten ranking for both gain ratio-based and model-based ranking methods. Lastly, we further reduce the feature set by eliminating highly correlated features.

Table 4.4: Top ten results from model-based and gain ratio-based ranking.

No.	Feature	Gain Ratio	Predictive Model
1	f_1		
2	f_2	✓	
3	f_3	✓	
4	f_4		
5	f_5		
6	f_6		
7	f_7		
8	f_8	✓	✓
9	f_9	✓	✓
10	f_{10}	✓	✓
11	f_{11}	✓	✓
12	f_{12}	✓	✓
13	f_{13}		
14	f_{14}		
15	f_{15}		
16	f_{16}		✓
17	f_{17}	✓	✓
18	f_{18}	✓	✓
19	f_{19}		
20	f_{20}		
21	f_{21}		
22	f_{22}		
23	f_{23}		
24	f_{24}		
25	f_{25}		
26	f_{26}		✓
27	f_{27}		
28	f_{28}		
29	f_{29}		
30	f_{30}	✓	✓

Table 4.5: Final feature subset.

No.	Feature	Description
1	F_{30}	Retinal Activation
2	f_{12}	Saccade Duration
3	f_{17}	Density Metric 1
4	f_{10}	Sum of absolute angles
5	f_{18}	Density Metric 2
6	f_8	Saccade Length

Table 4.6: Detailed performance metrics of predictive model for biometric identification using sketch-based features from eye movement.

Class	**Accuracy**	**F-Measure**	**ROC Area**	**ZeroR**
A1	0.94	0.87	0.989	0.1
E1	0.85	0.87	0.991	0.1
N1	0.93	0.94	0.992	0.1
E2	0.85	0.88	0.991	0.1
A2	0.83	0.87	0.987	0.1
A3	0.89	0.92	0.992	0.1
E3	0.83	0.84	0.987	0.1
N2	0.9	0.88	0.997	0.1
N3	0.91	0.9	0.997	0.1
A4	0.96	0.95	0.998	0.1
Weighted Avg.	0.89	0.89	0.99	0.1

4.4.2 Classification Results

We conducted performance tests using the sketch-based eye movement feature subset presented in Table 4.5 by developing a within-subject predictive model using a Random Forest classifier for each test case [37]. Each predictive model was evaluated using a k-fold cross-validation partitioning scheme ($k = 10$). Multiple (k) rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds in order to reduce variability. The aggregated results over all rounds serves as the performance evaluation of the predictive model, and a good approximation of how the model will perform in the real world. All training and testing evaluations were performed using WEKA software package [84], an open source machine learning software for building and testing predictive models. For comparison purposes, we include the results of a ZeroR classifier. The ZeroR classifier is a simple majority rule or mode rule classifier, which always returns the majority or modal class for all input test samples independent of the feature values of the input sample. The results of the ZeroR classifier is equivalent to random chance and serves

as baseline for both sketched-based and shapelet-based classifiers.

In Table 4.6, we report performance statistics (area under receiver operating characteristics (ROC) curve (AUC), accuracy, and f-measure) on predicting the identity of the image reader using eye movement features. The confusion matrix provided in Table 4.7 illustrates the instances of error when predicting the actual class label.

Table 4.7: Confusion matrix for predictive model using sketch-based features from eye movement.

N=1000	**PREDICTED CLASS**									
ACTUAL CLASS		A1	E1	N1	E2	A2	A3	E3	N2	N3	A4
	A1	94	0	1	1	0	0	2	2	0	0
	E1	0	85	1	7	0	0	3	2	2	0
	N1	2	0	93	0	0	1	3	1	0	0
	E2	3	8	0	85	2	0	0	1	1	0
	A2	2	2	2	2	83	3	5	0	1	0
	A3	1	0	4	0	3	89	0	3	0	0
	E3	4	2	3	3	1	0	83	1	3	0
	N2	1	4	0	0	1	0	0	90	1	3
	N3	3	0	2	0	0	0	1	0	91	3
	A4	2	1	0	0	0	0	0	1	0	96

Next, we applied the same cross-validation partitioning scheme, and conducted performance tests using sketch-based eye movement features to predict the experience level of each image reader. In Table 4.8, we report the performance statistics for predicting the image readers' experience level, and the corresponding confusion matrix in Table 4.9.

4.4.3 Comparison with Alternative Methods

In this section, we compare the performance of a biometric identification model using eye movement features, with an alternative shapelet-based feature set. To this

Table 4.8: Detailed performance metrics of predictive model for experience-level using shapelet-based features from pupillary changes.

Class	**Accuracy**	**F-Measure**	**ROC Area**	**ZeroR**
NR	0.90	0.9	0.99	0.4
AR	0.93	0.93	0.99	0.4
E	0.91	0.91	0.98	0.4
Weighted Avg.	0.91	0.91	0.99	0.4

Table 4.9: Confusion matrix of predictive model for experience level using sketch-based features from eye movement.

n = 1000	**PREDICTED CLASS**			
ACTUAL CLASS		NR	AR	E
	NR	270	17	13
	AR	15	370	15
	E	15	13	272

end, we performed a time-series shapelet analysis on percentage change in pupil dilation to identify shaplets that provide discriminative information about individual image readers. First, we developed a dictionary of 300 maximally informative shapelets of varied lengths ($1s - 3s$) using the methods presented in [299].

Next, using each unique shapelet as a search term, we computed the *term frequencyinverse document frequency* [156, 249] ($tf - idf$) score, a commonly used term weighting scheme in information retrieval systems, for each shapelet in the dictionary. Using this method, we represented each mammographic case reading as a vector of $tf - idf$ scores of length 300 (a score representing each shapelet). This method of representation using vectors in a uniform vector space is known as the vector space model and is fundamental to a host of information retrieval operations ranging from scoring documents on a query, document classification and document clustering [237].

Because the number of features was comparatively high (300) in proportion to

Table 4.10: Detailed performance metrics of predictive model for biometric identification using shapelet-based features from pupillary changes.

Class	**Accuracy**	**F-Measure**	**ROC Area**	**ZeroR**
A1	0.54	0.507	0.885	0.1
E1	0.61	0.663	0.941	0.1
N1	0.67	0.663	0.946	0.1
E2	0.37	0.363	0.822	0.1
A2	0.65	0.634	0.943	0.1
A3	0.97	0.937	0.997	0.1
E3	0.52	0.505	0.905	0.1
N2	0.53	0.541	0.937	0.1
N3	0.38	0.406	0.871	0.1
A4	0.46	0.469	0.869	0.1
Weighted Average	0.57	0.569	0.912	0.1

the number of data samples (total number of mammographic case readings is 1000), along with the associated increase in memory and computation costs, we performed dimensionality reduction (feature subset selection) for each image reader's data set. We used a wrapper feature subset evaluation method [137], which searches for an optimal subset of ten or fewer features by evaluating feature subsets using a learning scheme. The search function was performed using the simple genetic algorithm described by Goldberg in [78]. The accuracy of the learning scheme for each subset of features was estimated using a Random Forest [37] classifier with 10-fold cross-validation. The final set of features selected for use in the optimal subset were those features, which were selected in at least half of the ten folds in the wrapper subset evaluation. All training and testing evaluations were performed using the WEKA software package [84].

4.5 Discussion

Based on the ranking scores discussed in detail in Section 4.4.1, the final set of features (see Table 4.5) include four measures related to motion: the retinal activa-

Table 4.11: Confusion matrix of predictive model for biometric identification using shapelet-based features from pupillary changes.

N=1000		PREDICTED CLASS									
ACTUAL CLASS		A1	E1	N1	E2	A2	A3	E3	N2	N3	A4
	A1	54	0	7	1	9	2	4	13	0	10
	E1	0	61	0	14	6	0	9	0	8	2
	N1	10	0	67	0	1	3	0	15	0	4
	E2	7	10	0	37	9	0	11	1	18	7
	A2	7	4	0	9	65	0	3	6	4	2
	A3	0	0	1	0	0	97	0	1	0	1
	E3	6	5	0	14	3	0	52	0	11	9
	N2	13	0	23	0	2	3	0	53	0	6
	N3	2	2	0	24	10	0	15	0	38	9
	A4	14	2	4	5	0	2	12	7	8	46

tion, and the duration, length, and rotational change of the shape formed by the saccade, and two measures of visual appearance: ratio of saccade length to overall size (*f16*), and the ratio of saccade length to the actual inter-fixation distance (*f17*). The highest ranked feature, retinal activation, explains the tendency of the image readers' saccadic scanpath to follow a specific direction related to cortical maps. Cortical maps are collections (areas) of the brain identified as being responsible for processing a specific type of information. Neurons in the visual cortex are grouped together based on similar response properties that represent stimulus features such as edge orientation, direction of motion, and position in space. We speculate that the retinal activation feature is related to individual behavioral adaptations resulting in increased sensitivity of specific regions in the visual cortex. However, a more detailed study and experimental data is required to validate this speculative statement.

Previous studies in mammography have identified some measures of direction, duration, and lengths of saccadic movements as containing discriminative information about the experience of an image reader [150, 193, 143]. To the best of our knowledge,

these features were never applied in predictive models as biometric identifiers or for predicting experience level in mammography. Additionally, the characterization using gesture recognition methods have never been explored until now.

Intuitively, both density metrics (*f16* and *f17*) capture the spatial efficiency of the saccadic movement. While *f16* measures the linear efficiency of the scanpath, *f17* measures the two-dimensional spatial efficiency of the scanpath. Both features give a piecewise decomposition of the geometric properties of the scanpath formed by the image reader during the screening process. Previous studies have suggested that measures of overall scanpath formed during the viewing of a mammographic case are related to the individual and experience [150, 195]. The scanpath has also been studied as a biometric for individual identification under varied image viewing conditions unrelated to mammography [301, 114].

4.6 Conclusions

In this study, we proposed and evaluated two methods for extracting features, which contain meaningful information about the individual image reader and their level of expertise in screening mammography. These features characterize changes in positional and non-positional changes in the eyes. First, we applied sketch gesture recognition algorithms to extract geometric-based features from positional eye-movement data. These features give a fine-grained characterization of the scanpath by aggregating the spatial (shape), directional, and kinetic properties of its constituent saccadic movements. The second method applies timeseries shapelet analysis to extract discriminative information to from changes in pupil dilation from pupillometric data.

Using a corpus of eye movement and pupillary data from 100 mammographic cases reviewed by ten image readers recorded under clinically equivalent experimental

conditions, the findings presented in this study establishes the following generalizable trends:

1. During the mammographic screening task, positional and non-positional measures of changes in the eye can provide sufficient discriminative information about the identity of the image reader

2. Positional and non-positional measures of changes in the eye provide sufficient discriminative characterization of the image readers' experience level during mammographic screening.

3. Compared to non-positional measures (shapelet-based pupilary features), positional measures of change in the eye (sketch-based features) performed better at predicting the identity and experience level of the image reader.

4. Both positional and non-positional measures perform significantly better than random chance at predicting the image readers' identity and experience level.

5. DISCUSSION

In this work, we developed novel eye-event primitives and extended existing algorithms, inspired from other domains including sketch recognition, data mining, and information retrieval to improve prediction of mammographic case characteristics (such as case pathology and breast parenchyma density), radiologists characteristics (including individual identity and experience level), and risk of diagnostic error.

5.1 Fractal Dimension of Scanpath

We have developed an eye-event feature, fractal dimension, which requires no prior knowledge of regions of interest in stimulus image. The fractal dimension, which characterizes the space filling capacity of a pattern, provides a statistical index of complexity of radiologist's scanpath during mammographic screening. This index, we hypothesized, can accurately characterize the characteristics of a mammographic case, the image reader's identity and experience, and the risk of diagnostic error.

Based on results from our analysis of a corpus of eye-event data from 10 image readers with varied levels of experience and expertise, recorded during diagnostic screening of 100 digitized screen film mammographic images, comprising varied pathology and breast parenchyma density, our findings indicate that the application of fractal dimension to characterize the complexity of gaze scanpath, results in a new eye-event feature, which can be used in predictive modelling. The findings presented in this study support our hypothesis and establish the following generalizable trends:

1. The characteristics of a mammographic case (pathology, and breast parenchyma density) are independent factors in predicting complexity of visual search behavior.

2. The characteristics of the image reader (individual, and level of experience) are independent factors in predicting complexity of visual search behavior.

3. The pathology and breast parenchyma density of a mammographic case, experience level of the image reader, and the resulting diagnostic decision combine as predictors of complexity of visual search behavior during mammographic screening.

4. The visual search complexity while viewing cases with normal pathology are significantly different from cases with malignant pathology.

5. The visual search complexity increases monotonically with increasing breast parenchyma density of a mammographic image. Effectively, low-density mammographic images correspond to lower visual search complexity, while medium-density images correspond to a higher visual search complexity, and high-density images correspond to the highest visual search complexity. This finding is consistent with results obtained by Al Mousa et al. [185], who reported significant increases in visual search parameters when comparing low- and high-density mammograms.

6. On average, the visual search complexity of Radiology residents (both new trainees and advance trainees groups) is significantly lower than the average complexity of experienced radiologists.

7. There are notable differences in visual search complexity between individual radiologists.

5.2 Time Series Shapelet Analysis of Eye-Events

Prior findings show that pupillary changes capture information about what is perceived, cognitive/mental processing, and cognitive/mental load. Deducing from

these prior research findings, we tested the efficacy of pupillary features as predictors of task performance. Our initial analysis showed significant effects from eye movement features. These effects did not translate to high predictive performance on average using aggregate data primitives, which are typical in machine learning algorithms.

We implemented an existing machine learning algorithm, time series shapelet analysis, originally developed for the data mining and information retrieval research domain, and applied it for eye tracking research in mammography. In addition, we extended the time series shapelet analysis algorithm in a manner that optimizes its utility for eye tracking data. The shapelet analysis algorithm with our novel extensions, we hypothesized, more accurately characterize pupillary changes and result in better predictive models. This technique is generalizable and can be utilized for analyzing data in other domains.

Based on results from our analysis of a corpus of eye-event data from 10 image readers with varied levels of experience and expertise, recorded during diagnostic screening of 100 digitized screen film mammographic images, comprising varied pathology and breast parenchyma density, our findings indicate that the application of time series shapelet analysis to characterize a single type of eye-event (pupil dilation), achieves a significant improvement in predictive accuracy. The findings presented in this study support our hypothesis and establish the following generalizable trends

1. The characteristics of a mammographic case (pathology and breast parenchyma density) are independent factors in predicting eye movement and pupillary changes during mammographic screening.

2. The experience level of the image reader is a significant factor in predicting eye movement and pupillary changes during mammographic screening.

3. The performance of aggregate features from eye-movement and pupillary changes marginally outperform random chance at predicting the characteristics of a mammographic case and the image reader's diagnostic performance.

4. Measures such as time series shapelets, which capture temporal changes in eye movement and pupillary changes, perform significantly better than random chance at predicting the characteristics of a mammographic case and the image reader's diagnostic performance.

5.3 Gesture Recognition of Saccadic Eye Movements

In this study, we proposed and evaluated new techniques for extracting eye-event features, which contain meaningful information about the individual image reader and their level of expertise in screening mammography. First, we applied sketch gesture recognition algorithms to extract geometric-based features from eye-event data using a head-mounted eye tracking sensor. These features, we hypothesized, are capable of providing a more fine-grained characterization of the scanpath by aggregating the spatial (shape), directional, and kinetic properties of its constituent saccadic movements. We compared features described above with a second, previously discussed method, which applies timeseries shapelet analysis to extract discriminative information to from changes in pupil dilation from eye-event data.

Based on results from our analysis of a corpus of eye-event data from 10 image readers with varied levels of experience and expertise, recorded during diagnostic screening of 100 digitized screen film mammographic images, comprising varied pathology and breast parenchyma density, our findings indicate that saccadic movements can be accurately characterized using sketch gesture recognition algorithms. The findings presented in this study support our hypothesis and establish the following generalizable trends:

1. During the mammographic screening task, positional and non-positional measures of changes in the eye can provide sufficient discriminative information about the identity of the image reader

2. Positional and non-positional measures of changes in the eye provide sufficient discriminative characterization of the image readers' experience level during mammographic screening.

3. Compared to non-positional measures (shapelet-based pupilary features), positional measures of change in the eye (sketch-based features) performed better at predicting the identity and experience level of the image reader.

4. Both positional and non-positional measures perform significantly better than random chance at predicting the image readers' identity and experience level.

Generally we found distinct differences in the search strategies between the experienced and inexperienced image readers and we discussed the significance of these findings. We believe our overall results support some recent observations and theoretical models of expert performance. These findings may prove to be helpful for performance assessment in educational programmes both within the image interpretation for non-radiology practitioners and other domains.

REFERENCES

[1] Jans Aasman, Gijsbertus Mulder, and Lambertus JM Mulder. Operator effort and the measurement of heart-rate variability. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 29(2):161–170, 1987.

[2] Sylvia Ahern and Jackson Beatty. Pupillary responses during information processing vary with scholastic aptitude test scores. *Science*, 205(4412):1289–1292, 1979.

[3] Folami Alamudun, Jongyoon Choi, Ricardo Gutierrez-Osuna, Hira Khan, and Beena Ahmed. Removal of subject-dependent and activity-dependent variation in physiological measures of stress. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2012 6th International Conference on*, pages 115–122. IEEE, 2012.

[4] Folami Alamudun, Hong-Jun Yoon, Tracy Hammond, Kathy Hudson, Garnetta Morin-Ducote, and Georgia Tourassi. Shapelet analysis of pupil dilation for modeling visuo-cognitive behavior in screening mammography. In *Proc. SPIE*, volume 9787, pages 97870M–97870M–13, 2016.

[5] Folami T. Alamudun, Hong-Jun Yoon, Kathy Hudson, Garnetta Morin-Ducote, and Georgia Tourassi. Fractal analysis of radiologists' visual scanning pattern in screening mammography. In *Proc. SPIE*, volume 9416, pages 94160T–94160T–8, 2015.

[6] E Alberdi, A A Povyakalo, L Strigini, P Ayton, M Hartswood, R Procter, and R Slack. Use of computer-aided detection (cad) tools in screening mam-

mography: a multidisciplinary investigation. *The British Journal of Radiology*, 78(Spec No 1):S31–S40, 2005.

[7] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[8] E D Anderson, B B Muir, J S Walsh, and A E Kirkpatrick. The efficacy of double reading mammograms in breast screening. *Clinical Radiology*, 49(4):248–51, Apr 1994.

[9] Olufunmilola Atilola, Martin Field, Erin McTigue, Tracy Hammond, and Julie Linsey. Evaluation of a natural sketch interface for truss fbds and analysis. In *Frontiers in Education Conference (FIE), 2011*, pages S2E–1. IEEE, 2011.

[10] Olufunmilola Atilola, Martin Field, Erin McTigue, Tracy Hammond, and Julie Linsey. Mechanix: a sketch recognition truss tutoring system. In *ASME 2011 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pages 645–654. American Society of Mechanical Engineers, 2011.

[11] Olufunmilola Atilola, Stephanie Valentine, Hong-Hoe Kim, David Turner, Erin McTigue, Tracy Hammond, and Julie Linsey. Mechanix: A natural sketch interface tool for teaching truss analysis and free-body diagrams. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 28(02):169–192, 2014.

[12] Olufunmilola Atilola, Francisco Vides, Erin M Mctigue, Julie S Linsey, and Tracy Anne Hammond. Automatic identification of student misconceptions and errors for truss analysis. In *American Society for Engineering Education*. American Society for Engineering Education, 2012.

[13] Alan Baddeley. Working memory. *Science*, 255(5044):556–559, 1992.

[14] Brian P Bailey and Shamsi T Iqbal. Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 14(4):21, 2008.

[15] Corinne Balleyguier, Karen Kinkel, Jacques Fermanian, Sebastien Malan, Germaine Djen, Patrice Taourel, and Olivier Helenon. Computer-aided detection (cad) in mammography: does it help the junior or the senior radiologist? *European Journal of Radiology*, 54(1):90–96, 2005.

[16] James R Barrett, Samuel J Dwyer III, Michael B Merickel, Thomas E Hutchinson, et al. Unobtrusively tracking mammographers' eye gaze direction and pupil diameter. In *Medical Imaging 1994*, pages 46–54. International Society for Optics and Photonics, 1994.

[17] Joey Bartley, Jonathon Forsyth, Prachi Pendse, Da Xin, Garrett Brown, Paul Hagseth, Ashish Agrawal, Daniel W Goldberg, and Tracy Hammond. World of workout: a contextual mobile rpg to encourage long term fitness. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health*, pages 60–67. ACM, 2013.

[18] Andrew Baum and Donna M Posluszny. Health psychology: mapping biobehavioral contributions to health and illness. *Annual Review of Psychology*, 50(1):137–163, 1999.

[19] Craig A Beam. Interpretion error in mammography: taxonomy and measurement. *Seminars in Breast Disease*, 6(3):153 – 157, 2003. Improving and Monitoring Mammographic Interpretive Skills.

[20] Jackson Beatty. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, 91(2):276, 1982.

[21] Jackson Beatty and Brennis Lucero-Wagoner. The pupillary system. *Handbook of Psychophysiology*, 2:142–162, 2000.

[22] Roman Bednarik, Tomi Kinnunen, Andrei Mihaila, and Pasi Fränti. Eye-movements as a biometric. In *Image Analysis*, pages 780–789. Springer, 2005.

[23] D Beijerinck, JJ Deurenberg, JJ Rombach, and K Borsje. Breast cancer screening: all's well that ends well, or much ado about nothing? *American Journal of Roentgenology*, 152(4):891–891, 1989.

[24] Gershon Ben-Shakhar. Standardization within individuals: A simple method to neutralize individual differences in skin conductance. *Psychophysiology*, 22(3):292–299, 1985.

[25] L Berlin. Malpractice issues in radiology errors in judgment. *AJR. American Journal of Roentgenology*, 166(6):1259–1261, 1996.

[26] Leonard Berlin. Malpractice issues in radiology. perceptual errors. *AJR. American Journal of Roentgenology*, 167(3):587–590, 1996.

[27] Leonard Berlin. Accuracy of diagnostic procedures: has it improved over the past five decades? *AJR. American Journal of Roentgenology*, 188(5):1173–1178, 2007.

[28] Akshay Bhat and Tracy Hammond. Using entropy to distinguish shape versus text in hand-drawn diagrams. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 1395–1400. Morgan Kaufmann Publishers Inc., 2009.

[29] RE Bird. Low-cost screening mammography: report on finances and review of 21,716 consecutive cases. *Radiology*, 171(1):87–90, 1989.

[30] Richard E Bird, Terry W Wallace, and Bonnie C Yankaskas. Analysis of cancers missed at screening mammography. *Radiology*, 184(3):613–617, 1992.

[31] Robyn L Birdwell, Debra M Ikeda, Kathryn F OShaughnessy, and Edward A Sickles. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection 1. *Radiology*, 219(1):192–202, 2001.

[32] Archie Bleyer and H Gilbert Welch. Effect of three decades of screening mammography on breast-cancer incidence. *The New England journal of medicine*, 367(21):1998–2005, Nov 2012.

[33] Wolfram Boucsein. *Electrodermal Activity*. Springer Science & Business Media, 2012.

[34] Wolfram Boucsein, Andrea Haarmann, and Florian Schaefer. Combining skin conductance and heart rate variability for adaptive automation during simulated ifr flight. In *Engineering Psychology and Cognitive Ergonomics*, pages 639–647. Springer, 2007.

[35] Bruno Boyer, Corinne Balleyguier, Olivier Granat, and Christian Pharaboz. Cad in questions/answers: Review of the literature. *European Journal of Radiology*, 69(1):24–33, 2009.

[36] Margaret M Bradley, Laura Miccoli, Miguel A Escrig, and Peter J Lang. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–7, Jul 2008.

[37] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[38] Rachel F Brem, Janet Baum, Mary Lechner, Stuart Kaplan, Stuart Souders, L Gill Naul, and Jeff Hoffmeister. Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial. *American Journal of Roentgenology*, 181(3):687–693, 2003.

[39] Mireille Broeders, Sue Moss, Lennarth Nyström, Sisse Njor, Håkan Jonsson, Ellen Paap, Nathalie Massat, Stephen Duffy, Elsebeth Lynge, and Eugenio Paci. The impact of mammographic screening on breast cancer mortality in europe: a review of observational studies. *Journal of medical screening*, 19(suppl 1):14–25, 2012.

[40] MJM Broeders, NC Onland-Moret, HJTM Rijken, JHCL Hendriks, ALM Verbeek, and R Holland. Use of previous screening mammograms to identify features indicating cases that would have a possible gain in prognosis following earlier detection. *European Journal of Cancer*, 39(12):1770–1775, 2003.

[41] Jackie Brown, Stirling Bryan, Ruth Warren, et al. Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. *BMJ*, 312(7034):809–812, 1996.

[42] John T Cacioppo, Louis G Tassinary, and Gary Berntson. *Handbook of psychophysiology*. Cambridge University Press, 2007.

[43] Renato Campanini, Danilo Dongiovanni, Emiro Iampieri, Nico Lanconelli, Matteo Masotti, Giuseppe Palermo, Alessandro Riccardi, and Matteo Roffilli. A novel featureless approach to mass detection in digital mammograms based on support vector machines. *Physics in Medicine and Biology*, 49(6):961, 2004.

[44] Paola Casti, Arianna Mencattini, Marcello Salmeri, and Rangaraj M Rangayyan. Analysis of structural similarity in mammograms for detection of

bilateral asymmetry. *Medical Imaging, IEEE Transactions on*, 34(2):662–671, 2015.

[45] Michelene TH Chi, Robert Glaser, and Marshall J Farr. *The nature of expertise*. Psychology Press, 2014.

[46] Heeyoul Choi, Brandon Paulson, and Tracy Hammond. Gesture recognition based on manifold learning. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 247–256. Springer, 2008.

[47] Jae Young Choi, Dae Hoe Kim, Konstantinos N Plataniotis, and Yong Man Ro. Computer-aided detection (cad) of breast masses in mammography: combined detection and ensemble classification. *Physics in medicine and biology*, 59(14):3697, 2014.

[48] Jongyoon Choi, Beena Ahmed, and Ricardo Gutierrez-Osuna. Ambulatory stress monitoring with minimally-invasive wearable sensors. *Computer. Sci. and Eng*, 2010.

[49] Jongyoon Choi and Ricardo Gutierrez-Osuna. Using heart rate monitors to detect mental stress. In *Wearable and Implantable Body Sensor Networks, 2009. BSN 2009. 6th International Workshop on*, pages 219–223. IEEE, 2009.

[50] Jongyoon Choi and Ricardo Gutierrez-Osuna. Estimating mental stress using a wearable cardio-respiratory sensor. In *Sensors, 2010 IEEE*, pages 150–154. IEEE, 2010.

[51] Çağla Çığ and Tevfik Metin Sezgin. Gaze-based prediction of pen-based virtual interaction tasks. *International Journal of Human-Computer Studies*, 73:91 – 106, 2015.

[52] OpenStax College. The extrinsic eye muscles. https://commons.wikimedia.org/wiki/File:1107_The_Extrinsic_Eye_Muscles.jpg, May 2013.

[53] Wikimedia Commons. Schematic diagram of the human eye in english. https://commons.wikimedia.org/wiki/File:Schematic_diagram_of_the_human_eye_with_English_annotations.svg, Aug 2007.

[54] Andrew RA Conway. Individual differences in working memory capacity: More evidence for a general capacity theory. *Memory*, 4(6):577–590, 1996.

[55] Paul Corey and Tracy Hammond. Gladder: combining gesture and geometric sketch recognition. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, pages 1788–1789. AAAI Press, 2008.

[56] Danielle Cummmings, Francisco Vides, and Tracy Hammond. I don't believe my eyes!: geometric sketch recognition for a computer art tutorial. In *Proceedings of the International Symposium on Sketch-Based Interfaces and Modeling*, pages 97–106. Eurographics Association, 2012.

[57] Katie Dahmen and Tracy Hammond. Distinguishing between sketched scribble look alikes. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, pages 1790–1791. AAAI Press, 2008.

[58] Farzin Deravi and Shivanand P Guness. Gaze trajectory as a biometric modality. In *BIOSIGNALS*, pages 335–341, 2011.

[59] Daniel Dixon, Manoj Prasad, and Tracy Hammond. icandraw: using sketch recognition and corrective feedback to assist a user in drawing human faces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 897–906. ACM, 2010.

[60] Daniel Meyer Dixon. *A methodology using assistive sketch recognition for improving a persons ability to draw.* PhD thesis, Texas A&M University, 2009.

[61] K Doi, H MacMahon, S Katsuragawa, R M Nishikawa, and Y Jiang. Computer-aided diagnosis in radiology: potential and pitfalls. *European journal of radiology*, 31(2):97–109, Aug 1999.

[62] Tim Donovan, David J Manning, Peter W Phillips, Stephen Higham, and Trevor Crawford. The effect of feedback on performance in a fracture detection task. In *Medical Imaging*, pages 79–85. International Society of Photo Optical Instrumentation Engineers, 2005.

[63] Marie-Pierre Dubuisson and Anil K Jain. A modified hausdorff distance for object matching. In *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 566–568 vol.1, Oct 1994.

[64] Andrew Duchowski. *Eye tracking methodology: Theory and Practice*, volume 373. Springer Science & Business Media, 2007.

[65] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification.* John Wiley & Sons, 2012.

[66] Lucien EM Duijm, Johanna H Groenewoud, Jan HCL Hendriks, and Harry J de Koning. Independent double reading of screening mammograms in the netherlands: Effect of arbitration following reader disagreements 1. *Radiology*, 231(2):564–570, 2004.

[67] Joann G Elmore, Mary B Barton, Victoria M Moceri, Sarah Polk, Philip J Arena, and Suzanne W Fletcher. Ten-year risk of false positive screening mam-

mograms and clinical breast examinations. *New England Journal of Medicine*, 338(16):1089–1096, 1998.

[68] Brian David Eoff and Tracy Hammond. Who dotted that'i'?: context free user differentiation through pressure and tilt pen data. In *Proceedings of Graphics Interface 2009*, pages 149–156. Canadian Information Processing Society, 2009.

[69] Lenore I Everson. Mammographic interpretation: A practical approach. *Radiology*, 185(3):864–864, 1992.

[70] Torbjörn Falkmer, Joakim Dahlman, Tania Dukic, Anna Bjällmark, and Matilda Larsson. Fixation identification in centroid versus start-point modes using eye-tracking data. *Perceptual and Motor Skills*, 106(3):710–724, 2008.

[71] Tom Fearn. On orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems*, 50(1):47–52, 2000.

[72] Martin Field, Stephanie Valentine, Julie Linsey, and Tracy Hammond. Sketch recognition algorithms for comparing complex and unpredictable shapes. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence - Volume 3*, IJCAI'11, pages 2436–2441. AAAI Press, 2011.

[73] Michael Fligner, Joseph Verducci, Jeff Bjoraker, and Paul Blower. A new association coefficient for molecular dissimilarity. In *The 2nd Joint Sheffield Conference on Chemoinformatics*, 2001.

[74] Timothy W Freer and Michael J Ulissey. Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center 1. *Radiology*, 220(3):781–786, 2001.

[75] Chiara Galdi, Michele Nappi, Daniel Riccio, Virginio Cantoni, and Marco Porta. A new gaze analysis based soft-biometric. In *MCPR*, pages 136–144.

Springer, 2013.

[76] Kavita Ganesan, URajendra Acharya, Chua Kuang Chua, Lim Choo Min, K Thomas Abraham, and Kung Bo Ng. Computer-aided breast cancer detection using mammograms: a review. *Biomedical Engineering, IEEE Reviews in*, 6:77–98, 2013.

[77] Daniel W Goldberg, Myles G Cockburn, Tracy A Hammond, Geoffrey M Jacquez, Daniel Janies, Craig Knoblock, Werner Kuhn, Edward Pultar, and Martin Raubal. Envisioning a future for a spatial-health cybergis marketplace. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health*, pages 27–30. ACM, 2013.

[78] David E Goldberg et al. *Genetic Algorithms in Search, Optimization, and Machine Learning*, volume 412. Addison-wesley Reading Menlo Park, 1989.

[79] US Cancer Statistics Working Group et al. United states cancer statistics: 1999–2006 incidence and mortality web-based report. *Atlanta, GA*, 2013.

[80] US Cancer Statistics Working Group et al. United states cancer statistics: 1999–2011 incidence and mortality web-based report. *Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute*, 2014.

[81] Q Guo, J Shao, and V Ruiz. Investigation of support vector machine for the detection of architectural distortion in mammographic images. *Journal of Physics: Conference Series*, 15(1):88, 2005.

[82] Andrea Haarmann, Wolfram Boucsein, and Florian Schaefer. Combining electrodermal responses and cardiovascular measures for probing adaptive automation during simulated flight. *Applied Ergonomics*, 40(6):1026–1040, 2009.

[83] Daniel Hahnemann and Jackson Beatty. Pupillary responses in a pitch-discrimination task. *Perception & Psychophysics*, 2(3):101–105, 1967.

[84] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[85] Mark A Hall. *Correlation-Based Feature Selection for Machine Learning*. PhD thesis, The University of Waikato, 1999.

[86] Tracy Hammond. Natural sketch recognition in uml class diagrams. In *Proceedings of the MIT Student Oxygen Workshop*, 2001.

[87] Tracy Hammond. *Sketch Recognition: Algorithms and Applications*. Cambridge University Press, 2017. draft from March 1, 2016, publication forthcoming.

[88] Tracy Hammond and Randall Davis. Tahuti: A geometrical sketch recognition system for uml class diagrams. In *Proceedings of AAAI Spring Symposium on Sketch Understanding*. AAAI, 2002.

[89] Tracy Hammond and Randall Davis. Ladder: a language to describe drawing, display, and editing in sketch recognition. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 461–467. Morgan Kaufmann Publishers Inc., 2003.

[90] Tracy Hammond and Randall Davis. Automatically transforming symbolic shape descriptions for use in sketch recognition. In *Proceedings of the 19th National Conference on Artificial Intelligence*, pages 450–456. AAAI Press, 2004.

[91] Tracy Hammond and Randall Davis. Shady: A shape description debugger for use in sketch recognition. In *AAAI Fall Symposium on Making Pen-Based*

Interaction Intelligent and Natural, 2004.

[92] Tracy Hammond and Randall Davis. Ladder, a sketching language for user interface developers. *Computers & Graphics*, 29(4):518–532, 2005.

[93] Tracy Hammond and Randall Davis. Interactive learning of structural shape descriptions from automatically generated near-miss examples. In *Proceedings of the 11th International Conference on Intelligent user Interfaces*, pages 210–217. ACM, 2006.

[94] Tracy Hammond and Randall Davis. Tahuti: A geometrical sketch recognition system for uml class diagrams. In *ACM SIGGRAPH 2006 Courses*. ACM, 2006.

[95] Tracy Hammond and Randall Davis. Creating the perception-based ladder sketch recognition language. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems*, pages 141–150. ACM, 2010.

[96] Tracy Hammond, Brian Eoff, Brandon Paulson, Aaron Wolin, Katie Dahmen, Joshua Johnston, and Pankaj Rajan. Free-sketch recognition: Putting the chi in sketching. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '08, pages 3027–3032, New York, NY, USA, 2008. ACM.

[97] Tracy Hammond, Krzysztof Gajos, Randall Davis, and Howard E Shrobe. An agent-based system for capturing and indexing software design meetings. In *In Proceedings of the International Workshop on Agents in Design (WAID02*. Citeseer, 2002.

[98] Tracy Hammond, Drew Logsdon, Joshua Peschel, Joshua Johnston, Paul Taele, Aaron Wolin, and Brandon Paulson. A sketch recognition interface that recognizes hundreds of shapes in course-of-action diagrams. In *CHI '10 Extended*

Abstracts on Human Factors in Computing Systems, CHI EA '10, pages 4213–4218, New York, NY, USA, 2010. ACM.

[99] Tracy Hammond and Barry OSullivan. Recognizing freeform hand-sketched constraint network diagrams by combining geometry and context. In *Proceedings of the Eurographics Ireland*, 2007.

[100] Tracy Hammond and Brandon Paulson. Recognizing sketched multistroke primitives. *ACM Transactions on Intelligent Systems and Technology*, 1(1):4:1–4:34, October 2011.

[101] Tracy Hammond, Manoj Prasad, and Daniel Dixon. Art 101: learning to draw through sketch recognition. In *Smart Graphics*, pages 277–280. Springer, 2010.

[102] Tracy Hammond, Metin Sezgin, Olya Veselova, Aaron Adler, Michael Oltmans, Christine Alvarado, and Rebecca Hitchcock. Multi-domain sketch recognition. In *Proceedings of the 2nd Annual MIT Student Oxygen Workshop*, 2002.

[103] Tracy A. Hammond and Randall Davis. Recognizing interspersed sketches quickly. In *Proceedings of Graphics Interface 2009*, GI '09, pages 157–166, Toronto, Ont., Canada, Canada, 2009. Canadian Information Processing Society.

[104] Tracy Anne Hammond. *Ladder: A perceptually-based language to simplify sketch recognition user interface development*. PhD thesis, Massachusetts Institute of Technology, 2007.

[105] Tracy Anne Hammond, Drew Logsdon, Brandon Paulson, Joshua Johnston, Joshua Peschel, Aaron Wolin, and Paul Taele. A sketch recognition system for recognizing free-hand course of action diagrams. In *22nd Innovative Applications of Artificial Intelligence Conference on Artificial Intelligence*, 2010.

[106] Tracy Tracy Hammond Hammond. Enabling instructors to develop sketch recognition applications for the classroom. In *Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual*, pages S3J–11. IEEE, 2007.

[107] PA Hancock and MH Chignell. Adaptive control in human-machine systems. In *Human Factors Psychology*, pages 305–345. North-Holland Publishing Co., 1987.

[108] Felix Hausdorff. Dimension und äußeres maß. *Mathematische Annalen*, 79(1-2):157–179, 1918.

[109] Jennifer A Healey and Rosalind W Picard. Detecting stress during real-world driving tasks using physiological sensors. *Intelligent Transportation Systems, IEEE Transactions on*, 6(2):156–166, 2005.

[110] Michael Heath, Kevin Bowyer, Daniel Kopans, P Kegelmeyer Jr, Richard Moore, Kyong Chang, and S Munishkumaran. Current status of the digital database for screening mammography. In *Digital mammography*, pages 457–460. Springer, 1998.

[111] Robert R Henderson, Margaret M Bradley, and Peter J Lang. Modulation of the initial light reflex during affective picture viewing. *Psychophysiology*, 51(9):815–818, 2014.

[112] Eckhard H Hess and James M Polt. Pupil size in relation to mental activity during simple problem-solving. *Science*, 143(3611):1190–1192, 1964.

[113] Jennifer L Hocking Schuler and WILLIAM H O'BRIEN. Cardiovascular recovery from stress and hypertension risk factors: A meta-analytic review. *Psychophysiology*, 34(6):649–659, 1997.

[114] C. Holland and O.V. Komogortsev. Biometric identification via eye movement scanpaths in reading. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–8, Oct 2011.

[115] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press, 2011.

[116] H. Hse and A.R. Newton. Sketched symbol recognition using zernike moments. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 1, pages 367–370 Vol.1, Aug 2004.

[117] Linda L Humphrey, Mark Helfand, Benjamin K S Chan, and Steven H Woolf. Breast cancer screening: a summary of the evidence for the u.s. preventive services task force. *Annals of internal medicine*, 137(5 Part 1):347–60, Sep 2002.

[118] Peter Ingwersen. Cognitive perspectives of information retrieval interaction: elements of a cognitive ir theory. *Journal of Documentation*, 52(1):3–50, 1996.

[119] National Cancer Institute. Mammograms fact sheet. `http://www.cancer.gov/types/breast/mammograms-fact-sheet`, March 2014. (Visited on 10/02/2015).

[120] Shamsi T Iqbal, Xianjun Sam Zheng, and Brian P Bailey. Task-evoked pupillary response to mental workload in human-computer interaction. In *CHI'04 Extended Abstracts on Human Factors in Computing Systems*, pages 1477–1480. ACM, 2004.

[121] Wilfrid Jänig. Autonomic nervous system. In *Human Physiology*, pages 333–370. Springer, 1989.

[122] Chun-Chu Jen and Shyr-Shen Yu. Automatic detection of abnormal mammograms in mammographic images. *Expert Systems with Applications*, 42(6):3048–3055, 2015.

[123] Arthur R Jensen and William D Rohwer. The stroop color-word test: A review. *Acta Psychologica*, 25:36–93, 1966.

[124] Joshua Johnston and Tracy Hammond. Computing confidence values for geometric constraints for use in sketch recognition. In *Proceedings of the 7th Sketch-Based Interfaces and Modeling Symposium*, pages 71–78. Eurographics Association, 2010.

[125] Daniel Kahneman. *Attention and Effort*. Citeseer, 1973.

[126] Cornelia Kappeler-Setz, Franz Gravenhorst, Johannes Schumm, Bert Arnrich, and Gerhard Tröster. Towards long term monitoring of electrodermal activity in daily life. *Personal and Ubiquitous Computing*, 17(2):261–271, 2013.

[127] Levent Burak Kara and Thomas F. Stahovich. An image-based, trainable symbol recognizer for hand-drawn sketches. *Computers & Graphics*, 29(4):501 – 517, 2005.

[128] Anna N Karahaliou, Ioannis S Boniatis, Spyros G Skiadopoulos, Filippos N Sakellaropoulos, Nikolaos S Arikidis, Eleni A Likaki, George S Panayiotakis, and Lena I Costaridou. Breast cancer diagnosis: analyzing texture of tissue surrounding microcalcifications. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, 12(6):731–8, Nov 2008.

[129] Nico Karssemeijer, Andrea Hupse, Maurice Samulski, Michiel Kallenberg, Carla Boetes, and Gerard den Heeten. An interactive computer aided de-

cision support system for detection of masses in mammograms. In *Digital Mammography*, pages 273–278. Springer, 2008.

[130] Pawel Kasprowski and Józef Ober. *Biometric Authentication: ECCV 2004 International Workshop, BioAW 2004, Prague, Czech Republic, May 15th, 2004. Proceedings*, chapter Eye Movements in Biometrics, pages 248–258. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[131] Brandon L Kaster, Emily R Jacobson, and Tracy A Hammond. Sssousa: Automatically generating secure and searchable data collection studies. In *International Workshop on Visual Languages and Computing. Redwood City, CA, USA: VLC*, 2009.

[132] Christos D Katsis, Nikolaos Katertsidis, George Ganiatsas, and Dimitrios I Fotiadis. Toward emotion recognition in car-racing drivers: A biosignal processing approach. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 38(3):502–512, 2008.

[133] Kourtney Kebodeaux, Martin Field, and Tracy Hammond. Defining precise measurements with sketched annotations. In *Proceedings of the Eighth Eurographics Symposium on Sketch-Based Interfaces and Modeling*, pages 79–86. ACM, 2011.

[134] Karla Kerlikowske, Deborah Grady, Susan M Rubin, Christian Sandrock, and Virginia L Ernster. Efficacy of screening mammography: a meta-analysis. *Jama*, 273(2):149–154, 1995.

[135] Hong-hoe Kim, Paul Taele, Stephanie Valentine, Erin McTigue, and Tracy Hammond. Kimchi: a sketch-based developmental skill classifier to enhance pen-driven educational interfaces for children. In *Proceedings of the Inter-*

national Symposium on Sketch-Based Interfaces and Modeling, pages 33–42. ACM, 2013.

[136] ANDERS M. KNUTZEN and JOHN J. GISVOLD. Likelihood of malignant disease for various categories of mammographically detected, nonpalpable breast lesions. *Mayo Clinic Proceedings*, 68(5):454 – 460, 1993.

[137] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1):273–324, 1997.

[138] Thomas M Kolb, Jacob Lichy, and Jeffrey H Newhouse. Comparison of the performance of screening mammography, physical examination, and breast us and evaluation of factors that influence them: An analysis of 27,825 patient evaluations 1. *Radiology*, 225(1):165–175, 2002.

[139] Elizabeth A Krupinski. Influence of experience on scanning strategies in mammography. In *Medical Imaging 1996*, pages 95–101. International Society for Optics and Photonics, 1996.

[140] Elizabeth A Krupinski. Current perspectives in medical image perception. *Attention, Perception, & Psychophysics*, 72(5):1205–1217, 2010.

[141] Elizabeth A. Krupinski, Anna R. Graham, and Ronald S. Weinstein. Characterizing the development of visual search expertise in pathology residents viewing whole slide images. *Human Pathology*, 44(3):357 – 364, 2013.

[142] Elizabeth A Krupinski, Harold L Kundel, Philip F Judy, and Calvin F Nodine. The medical image perception society. key issues for image perception research. *Radiology*, 209(3):611–612, 1998.

[143] Elizabeth A Krupinski, Allison A. Tillack, Lynne Richter, Jeffrey T. Henderson, Achyut K. Bhattacharyya, Katherine M. Scott, Anna R. Graham, Michael R.

Descour, John R. Davis, and Ronald S. Weinstein. Eye-movement study and human performance using telepathology virtual slides. implications for medical education and differences with experience. *Human Pathology*, 37(12):1543 – 1556, 2006.

[144] H L Kundel. Medical image perception. *Academic Radiology*, 2 Suppl 2:S108–10, Sep 1995.

[145] Harold L Kundel. How to minimize perceptual error and maximize expertise in medical imaging. In *Proc. SPIE*, volume 6515, pages 651508–1, 2007.

[146] Harold L Kundel and D John Wright. The influence of prior knowledge on visual search strategies during the viewing of chest radioqraphs 1. *Radiology*, 93(2):315–320, 1969.

[147] Harold L Kundel, Calvin F Nodine, and Dennis Carmody. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative Radiology*, 13(3):175–181, 1978.

[148] Harold L Kundel, Calvin F Nodine, Emily F Conant, and Susan P Weinstein. Holistic component of image perception in mammogram interpretation: gaze-tracking study 1. *Radiology*, 242(2):396–402, 2007.

[149] Harold L Kundel, Calvin F Nodine, and Elizabeth A Krupinski. Searching for lung nodules: visual dwell indicates locations of false-positive and false-negative decisions. *Investigative Radiology*, 24(6):472–478, 1989.

[150] Harold L. Kundel and Jr. Paul S. La Follette. Visual search patterns and experience with radiological images. *Radiology*, 103(3):523–528, 1972. PMID: 5022947.

[151] Wenzhe Li and Tracy Hammond. Using scribble gestures to enhance editing behaviors of sketch recognition systems. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '12, pages 2213–2218, New York, NY, USA, 2012. ACM.

[152] Wenzhe Li and Tracy Anne Hammond. Recognizing text through sound alone. In *25th AAAI Conference on Artificial Intelligence*, 2011.

[153] A. Chris Long, Jr., James A. Landay, Lawrence A. Rowe, and Joseph Michiels. Visual similarity of pen gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, pages 360–367, New York, NY, USA, 2000. ACM.

[154] Otto Lowenstein and Irene E Loewenfeld. The pupil. *The Eye*, 3:231–267, 1962.

[155] George Lucchese, Martin Field, Jimmy Ho, Ricardo Gutierrez-Osuna, and Tracy Hammond. Gesturecommander: Continuous touch-based gesture prediction. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '12, pages 1925–1930, New York, NY, USA, 2012. ACM.

[156] Hans Peter Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317, 1957.

[157] Patrick J. Lynch. Eye movements abductors. `https://en.wikipedia.org/wiki/File:Eye_movements_abductors.jpg`, Dec 2006.

[158] Patrick J. Lynch. Eye movements adductors. `https://en.wikipedia.org/wiki/File:Eye_movements_adductors.jpg`, Dec 2006.

[159] Patrick J. Lynch. Eye movements depressors. https://en.wikipedia.org/wiki/File:Eye_movements_depressors.jpg, Dec 2006.

[160] Patrick J. Lynch. Eye movements elevators. https://en.wikipedia.org/wiki/File:Eye_movements_elevators.jpg, Dec 2006.

[161] Patrick J. Lynch. Eye movements lateral rotation. https://en.wikipedia.org/wiki/File:Eye_movements_lateral_rot.jpg, Dec 2006.

[162] Patrick J. Lynch. Eye movements medial rotation. https://en.wikipedia.org/wiki/File:Eye_movements_medial.jpg, Dec 2006.

[163] Päivi Majaranta. *Gaze Interaction and Applications of Eye Tracking: Advances in Assistive Technologies: Advances in Assistive Technologies*. IGI Global, 2011.

[164] David Manning, Susan Ethell, Tim Donovan, and Trevor Crawford. How do radiologists do it? The influence of experience and training on searching for chest nodules. *Radiography*, 12(2):134 – 142, 2006.

[165] DJ Manning, SC Ethell, and Tim Donovan. Detection or decision errors? missed lung cancer from the posteroanterior chest radiograph. *Academic Radiology*, 2014.

[166] Angela B. Mariotto, Anne-Michelle Noone, Nadia Howlader, Hyunsoon Cho, Gretchen E. Keel, Jessica Garshell, Steven Woloshin, and Lisa M. Schwartz. Cancer survival: An overview of measures, uses, and interpretation. *JNCI Monographs*, 2014(49):145–186, 2014.

[167] Sandra P Marshall. The index of cognitive activity: Measuring cognitive workload. In *Human Factors and Power Plants, 2002. Proceedings of the 2002 IEEE 7th Conference on*, pages 7–5. IEEE, 2002.

[168] J E Martin, M Moskowitz, and J R Milbrath. Breast cancer missed by mammography. *AJR. American Journal of Roentgenology*, 132(5):737–9, May 1979.

[169] Susana Martinez-Conde, Stephen L Macknik, and David H Hubel. The role of fixational eye movements in visual perception. *Nature Reviews Neuroscience*, 5(3):229–240, 2004.

[170] James G May, Robert S Kennedy, Mary C Williams, William P Dunlap, and Julie R Brannan. Eye movement indices of mental workload. *Acta Psychologica*, 75(1):75–89, 1990.

[171] Maciej A. Mazurowski, Huiman X. Barnhart, Jay A. Baker, and Georgia D. Tourassi. Identifying error-making patterns in assessment of mammographic birads descriptors among radiology residents using statistical pattern recognition. *Academic Radiology*, 19(7):865 – 871, 2012.

[172] Maciej A. Mazurowski, Piotr A. Habas, Jacek M. Zurada, Joseph Y. Lo, Jay A. Baker, and Georgia D. Tourassi. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, 21(23):427 – 436, 2008. Advances in Neural Networks Research: {IJCNN} 072007 International Joint Conference on Neural Networks {IJCNN} 07.

[173] Bruce S McEwen and Eliot Stellar. Stress and the individual: mechanisms leading to disease. *Archives of Internal Medicine*, 153(18):2093–2101, 1993.

[174] Kristin J McLoughlin, Philip J Bones, and Nico Karssemeijer. Noise equalization for detection of microcalcification clusters in direct digital mammogram images. *Medical Imaging, IEEE Transactions on*, 23(3):313–320, 2004.

[175] Claudia Mello-Thoms. How does the perception of a lesion influence visual search strategy in mammogram reading? *Academic Radiology*, 13(3):275–288, 2006.

[176] Claudia Mello-Thoms. Visual search characteristics in mammogram reading: Sfm vs. ffdm. In *Proc. SPIE*, pages 76270D–76270D, 2010.

[177] Claudia Mello-Thoms, Stanley Dunn, Calvin F Nodine, Harold L Kundel, and Susan P Weinstein. The perception of breast cancer: what differentiates missed from reported cancers in mammography? *Academic Radiology*, 9(9):1004–1012, 2002.

[178] Claudia Mello-Thoms, Stanley M Dunn, Calvin F Nodine, and Harold L Kundel. The perception of breast cancers-a spatial frequency analysis of what differentiates missed from reported cancers. *Medical Imaging, IEEE Transactions on*, 22(10):1297–1306, 2003.

[179] Jace Miller and Tracy Hammond. Wiiolin: a virtual instrument using the wii remote. In *Proceedings of the 2010 Conference on New Interfaces for Musical Expression (NIME)*, pages 497–500, 2010.

[180] Peter Miller and Susan M Astley. Automated detection of breast asymmetries. In *BMVC*, pages 1–10, 1993.

[181] Peter Miller and Susan M Astley. Detection of breast asymmetry using anatomical features. In *IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology*, pages 433–442. International Society for Optics and Photonics, 1993.

[182] Luqman Mahmood Mina and Nor Ashidi Mat Isa. A review of computer-aided detection and diagnosis of breast cancer in digital mammography. *Journal of*

Medical Sciences, 15(3):110, 2015.

[183] Marilyn J Morton, Dana H Whaley, Kathleen R Brandt, and Kimberly K Amrami. Screening mammograms: Interpretation with computer-aided detectionprospective evaluation 1. *Radiology*, 239(2):375–383, 2006.

[184] SM Moss, Lennarth Nyström, Håkan Jonsson, E Paci, E Lynge, S Njor, and M Broeders. The impact of mammographic screening on breast cancer mortality in europe: a review of trend studies. *Journal of medical screening*, 19(suppl 1):26–32, 2012.

[185] Dana S. AL Mousa, Patrick C. Brennan, Elaine A. Ryan, Warwick B. Lee, Jennifer Tan, and Claudia Mello-Thoms. How mammographic breast density affects radiologists' visual search patterns. *Academic Radiology*, 21(11):1386 – 1393, 2014.

[186] Mark D Mugglestone, Alastair G Gale, Helen C Cowley, and ARM Wilson. Defining the perceptual processes involved with mammographic diagnostic errors. In *Medical Imaging 1996*, pages 71–77. International Society for Optics and Photonics, 1996.

[187] SA Narod, P Sun, C Wall, C Baines, and AB Miller. Impact of screening mammography on mortality from breast cancer before age 60 in women 40 to 49 years of age. *Current Oncology*, 21(5):217, 2014.

[188] Trevor Nelligan, Seth Polsley, Jaideep Ray, Michael Helms, Julie Linsey, and Tracy Hammond. Mechanix: A sketch-based educational interface. In *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion*, pages 53–56. ACM, 2015.

[189] KH Ng and M Muttarak. Advances in mammography have improved early detection of breast cancer. *Journal of Hong Kong College of Radiologists*, 6:126–131, 2003.

[190] Peter Nickel and Friedhelm Nachreiner. Sensitivity and diagnosticity of the 0.1-hz component of heart rate variability as an indicator of mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 45(4):575–590, 2003.

[191] Robert M Nishikawa, Yulei Jiang, Maryellen L Giger, Carl J Vyborny, Robert Schmidt, et al. Computer-aided detection of clustered microcalcifications. In *Systems, Man and Cybernetics, 1992., IEEE International Conference on*, pages 1375–1378. IEEE, 1992.

[192] Calvin F Nodine, Harold L Kundel, Sherri C Lauver, and Lawrence C Toto. Nature of expertise in searching mammograms for breast masses. In *Medical Imaging 1996*, pages 89–94. International Society for Optics and Photonics, 1996.

[193] Calvin F. Nodine, Harold L. Kundel, Claudia Mello-Thoms, Susan P. Weinstein, Susan G. Orel, Daniel C. Sullivan, and Emily F. Conant. How experience and training influence mammography expertise. *Academic Radiology*, 6(10):575 – 585, 1999.

[194] CF Nodine and HL Kundel. A visual dwell algorithm can aid search and recognition of missed lung nodules in chest radiographs. *Visual Search*, 2:399–406, 1990.

[195] David Noton and Lawrence Stark. Scanpaths in eye movements during pattern perception. *Science*, 171(3968):308–311, 1971.

[196] Robert D. O'Donnell and F. Thomas Eggemeier. *Workload assessment methodology*, pages 1–49. John Wiley & Sons, Oxford, England, 1986.

[197] American College of Radiology, BI-RADS Committee, et al. *ACR BI-RADS breast imaging and reporting data system: breast imaging atlas*. American College of Radiology, 2003.

[198] American College of Radiology. BI-RADS Committee and American College of Radiology. *Breast imaging reporting and data system*. American College of Radiology, 1998.

[199] Mette S Olufsen, Johnny T Ottesen, Hien T Tran, Laura M Ellwein, Lewis A Lipsitz, and Vera Novak. Blood pressure and blood flow variation during postural change from sitting to standing: model development and validation. *Journal of Applied Physiology*, 99(4):1523–1537, 2005.

[200] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[201] Tom Y. Ouyang and Randall Davis. A visual approach to sketched symbol recognition. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, IJCAI'09, pages 1463–1468, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

[202] Rabi Narayan Panda, Bijay Ketan Panigrahi, and Manas Ranjan Patro. Feature extraction for classification of microcalcifications and mass lesions in mammograms. *IJCSNS International Journal of Computer Science and Network Security*, 9(5):255–265, 2009.

[203] Sungmee Park and Sundaresan Jayaraman. Enhancing the quality of life through wearable technology. *Engineering in Medicine and Biology Magazine,*

IEEE, 22(3):41–48, 2003.

[204] B Paulson, A Wolin, J Johnston, and T Hammond. Sousa: sketch-based online user study applet. In *Proceedings of the 5th Eurographics Conference on Sketch-Based Interfaces and Modeling*, pages 81–88. Eurographics Association, 2008.

[205] Brandon Paulson, Danielle Cummings, and Tracy Hammond. Object interaction detection using hand posture cues in an office setting. *International Journal of Human-Computer Studies*, 69(1):19–29, 2011.

[206] Brandon Paulson, Brian Eoff, Aaron Wolin, Joshua Johnston, and Tracy Hammond. Sketch-based educational games: "drawin" kids away from traditional interfaces. In *Proceedings of the 7th International Conference on Interaction Design and Children*, IDC '08, pages 133–136, New York, NY, USA, 2008. ACM.

[207] Brandon Paulson and Tracy Hammond. A system for recognizing and beautifying low-level sketch shapes using ndde and dcr. In *20th Annual ACM Symposium on User Interface Software and Technology Posters*, 2007.

[208] Brandon Paulson and Tracy Hammond. Marqs: retrieving sketches learned from a single example using a dual-classifier. *Journal on Multimodal User Interfaces*, 2(1):3–11, 2008.

[209] Brandon Paulson and Tracy Hammond. Office activity recognition using hand posture cues. In *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction - Volume 2*, pages 75–78. British Computer Society, 2008.

[210] Brandon Paulson and Tracy Hammond. Paleosketch: Accurate primitive sketch recognition and beautification. In *Proceedings of the 13th International Confer-*

ence on Intelligent User Interfaces, IUI '08, pages 1–10, New York, NY, USA, 2008. ACM.

[211] Brandon Paulson, Pankaj Rajan, Pedro Davalos, Ricardo Gutierrez-Osuna, and Tracy Hammond. What!?! no rubine features?: Using geometric-based features to produce normalized confidence values for sketch recognition. In *HCC Workshop: Sketch Tools for Diagramming (VL/HCC)*, pages 57–63, 2008.

[212] Brandon Chase Paulson. *Rethinking pen input interaction: enabling freehand sketching through improved primitive recognition*. PhD thesis, Texas A&M University, 2010.

[213] Ronald K Pearson. Outliers in process modeling and identification. *Control Systems Technology, IEEE Transactions on*, 10(1):55–63, 2002.

[214] L Pescarini and I Inches. Systematic approach to human error in radiology. *La radiologia medica*, 111(2):252–267, 2006.

[215] Joshua M Peschel and Tracy Anne Hammond. Strat: a sketched-truss recognition and analysis tool. In *2008 International Workshop on Visual Languages and Computing (VLC) at the 14th International Conference on Distributed Multimedia Systems (DMS)*. Knowledge Systems Instistute, 2008.

[216] Joshua M Peschel, Brandon Paulson, and Tracy Hammond. A surfaceless pen-based interface. In *Proceedings of the 7th ACM Conference on Creativity and Cognition*, pages 433–434. ACM, 2009.

[217] Nicholas Petrick, Berkman Sahiner, Samuel G. Armato, Alberto Bert, Loredana Correale, Silvia Delsanto, Matthew T. Freedman, David Fryd, David Gur, Lubomir Hadjiiski, Zhimin Huo, Yulei Jiang, Lia Morra, Sophie Paquerault, Vikas Raykar, Frank Samuelson, Ronald M. Summers, Georgia Tourassi, Hi-

royuki Yoshida, Bin Zheng, Chuan Zhou, and Heang-Ping Chan. Evaluation of computer-aided detection and diagnosis systemsa). *Medical Physics*, 40(8), 2013.

[218] Rosalind W Picard and Jocelyn Scheirer. The galvactivator: A glove that senses and communicates skin conductivity. In *Proceedings 9th Int. Conf. on HCI*, 2001.

[219] Kurt Plarre, Andrew Raij, Syed Monowar Hossain, Amin Ahsan Ali, Motohiro Nakajima, Mustafa al'Absi, Emre Ertin, Thomas Kamarck, Santosh Kumar, Marcia Scott, et al. Continuous inference of psychological stress from sensory measurements collected in the natural environment. In *Information Processing in Sensor Networks (IPSN), 2011 10th International Conference on*, pages 97–108. IEEE, 2011.

[220] Beryl Plimmer and Isaac Freeman. A toolkit approach to sketched diagram recognition. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...But Not As We Know It - Volume 1*, BCS-HCI '07, pages 205–213, Swinton, UK, UK, 2007. British Computer Society.

[221] Ming-Zher Poh, Nicholas C Swenson, and Rosalind W Picard. A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *Biomedical Engineering, IEEE Transactions on*, 57(5):1243–1252, 2010.

[222] S. Pramanik and D. Bhattacharjee. Geometric feature based face-sketch recognition. In *Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conference on*, pages 409–415, March 2012.

[223] Manoj Prasad and Tracy Hammond. Observational study on teaching artifacts created using tablet pc. In *CHI '12 Extended Abstracts on Human Factors in*

Computing Systems, CHI EA '12, pages 301–316, New York, NY, USA, 2012. ACM.

[224] Dale Purves, George J Augustine, David Fitzpatrick, Lawrence C Katz, Anthony-Samuel Lamantia, James O McNamara, and S Mark Williams. *Neuroscience. 2nd*. Sunderland. MA: Sinauer Associates, 2001.

[225] Elizabeth A Rafferty, Jeong Mi Park, Liane E Philpotts, Steven P Poplack, Jules H Sumkin, Elkan F Halpern, and Loren T Niklason. Diagnostic accuracy and recall rates for digital mammography and digital mammography combined with one-view and two-view tomosynthesis: results of an enriched reader study. *AJR. American Journal of Roentgenology*, 202(2):273–281, 2014.

[226] Pankaj Rajan and T Hammond. From paper to machine: extracting strokes from images for use in sketch recognition. In *Proceedings of the 5th Eurographics Conference on Sketch-Based Interfaces and Modeling*, pages 41–48. Eurographics Association, 2008.

[227] Suraj Rajan. Horner's syndrome and autonomic innervation of the eye. `https://commons.wikimedia.org/wiki/File:Horner's_Syndrome_and_Autonomic_innervation_of_the_eye.svg`, March 2012.

[228] Vijay Rajanna, Folami Alamudun, Daniel Goldberg, and Tracy Hammond. Let me relax: Toward automated sedentary state recognition and ubiquitous mental wellness solutions. In *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare*, MOBIHEALTH'15, pages 28–33, ICST, Brussels, Belgium, Belgium, 2015. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[229] Rangaraj M Rangayyan, Fabio J Ayres, and JE Leo Desautels. A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs.

Journal of the Franklin Institute, 344(3):312–348, 2007.

[230] Rangaraj M Rangayyan, Liang Shen, Yiping Shen, JE Leo Desautels, Heather Bryant, Timothy J Terry, Natalka Horeczko, and M Sarah Rose. Improvement of sensitivity of breast cancer diagnosis with adaptive neighborhood contrast enhancement of mammograms. *Information Technology in Biomedicine, IEEE Transactions on*, 1(3):161–170, 1997.

[231] Vijay M Rao, David C Levin, Laurence Parker, Barbara Cavanaugh, Andrea J Frangos, and Jonathan H Sunshine. How widely is computer-aided detection used in screening and diagnostic mammography? *Journal of the American College of Radiology*, 7(10):802–805, 2010.

[232] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372, 1998.

[233] David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011.

[234] Dean Rubine. Specifying gestures by example. *SIGGRAPH Comput. Graph.*, 25(4):329–337, July 1991.

[235] William Rucklidge. *Efficient Visual Recognition Using the Hausdorff Distance*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996.

[236] Lizawati Salahuddin, Jaegeol Cho, Myeong Gi Jeong, and Desok Kim. Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. In *Engineering in Medicine and Biology Society, 2007. EMBS*

2007. 29th Annual International Conference of the IEEE, pages 4656–4659. IEEE, 2007.

[237] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[238] Maurice Samulski, Rianne Hupse, Carla Boetes, Roel DM Mus, Gerard J den Heeten, and Nico Karssemeijer. Using computer-aided detection in mammography as a decision support. *European radiology*, 20(10):2323–2330, 2010.

[239] K Sato and F Sato. Individual variations in structure and function of human eccrine sweat gland. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 245(2):R203–R208, 1983.

[240] RA Schmidt and CE Metz. Please be specific. *AJR. American journal of roentgenology*, 154(5):1121–1122, 1990.

[241] Brian S. Schnitzer and Eileen Kowler. Eye movements during multiple readings of the same text. *Vision Research*, 46(10):1611 – 1632, 2006.

[242] Johannes Schumm, Marc Bachlin, Cornelia Setz, Bert Arnrich, Daniel Roggen, and Gerhard Troster. Effect of movements on the electrodermal response after a startle event. In *2nd International Conference on Pervasive Computing Technologies for Healthcare*, pages 315–318. IEEE, 2008.

[243] Diane Scutt, Gillian A Lancaster, and John T Manning. Breast asymmetry and predisposition to breast cancer. *Breast Cancer Research*, 8(2):R14, 2006.

[244] Cornelia Setz, Bert Arnrich, Johannes Schumm, Roberto La Marca, G Troster, and Ulrike Ehlert. Discriminating stress from cognitive load using a wearable eda device. *Information Technology in Biomedicine, IEEE Transactions on*, 14(2):410–417, 2010.

[245] Mehmet Sezgin et al. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging*, 13(1):146–168, 2004.

[246] Nabeel Shahzad, Brandon Paulson, and Tracy Hammond. Urdu qaeda: recognition system for isolated urdu characters. In *Proceedings of the IUI Workshop on Sketch Recognition, Sanibel Island, Florida*, 2009.

[247] Robert A Smith, Vilma Cokkinides, Andrew C von Eschenbach, Bernard Levin, Carmel Cohen, Carolyn D Runowicz, Stephen Sener, Debbie Saslow, and Harmon J Eyre. American cancer society guidelines for the early detection of cancer. *CA: A Cancer Journal for Clinicians*, 52(1):8–22, 2002.

[248] American Cancer Society. *Cancer Facts & Figures*. American Cancer Society Atlanta, 1998.

[249] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[250] Janani C Sriram, Minho Shin, Tanzeem Choudhury, and David Kotz. Activity-aware ecg-based patient authentication for remote health monitoring. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, pages 297–304. ACM, 2009.

[251] Michael Sung. *Non-invasive wearable sensing systems for continuous health monitoring and long-term behavior modeling*. PhD thesis, Massachusetts Institute of Technology, 2005.

[252] Gbor J. Szkely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *Ann. Statist.*, 35(6):2769–2794, 12 2007.

[253] László Tabár, Bedrich Vitak, Hsiu-Hsi Tony Chen, Ming-Fang Yen, Stephen W Duffy, and Robert A Smith. Beyond randomized controlled trials. *Cancer*, 91(9):1724–1731, 2001.

[254] Paul Taele, Laura Barreto, and Tracy Hammond. Maestoso: An intelligent educational sketching tool for learning music theory. In *27th Innovative Applications of Artificial Intelligence Conference on Artificial Intelligence*, 2015.

[255] Paul Taele and Tracy Hammond. A geometric-based sketch recognition approach for handwritten mandarin phonetic symbols i. In *2008 International Workshop on Visual Languages and Computing (VLC) at the 14th International Conference on Distributed Multimedia Systems (DMS)*, 2008.

[256] Paul Taele and Tracy Hammond. Using a geometric-based sketch recognition approach to sketch chinese radicals. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, pages 1832–1833. AAAI Press, 2008.

[257] Paul Taele and Tracy Hammond. Hashigo: A next-generation sketch interactive system for japanese kanji. In *21st Innovative Applications of Artificial Intelligence Conference on Artificial Intelligence (IAAI)*, 2009.

[258] Paul Taele and Tracy Hammond. Lamps: A sketch recognition-based teaching tool for mandarin phonetic symbols i. *Journal of Visual Languages and Computing*, 21(2):109–120, 2010.

[259] Paul Taele and Tracy Hammond. Initial approaches for extending sketch recognition to beyond-surface environments. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '12, pages 2039–2044, New York, NY, USA, 2012. ACM.

[260] Paul Taele and Tracy Hammond. Adapting surface sketch recognition techniques for surfaceless sketches. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 3243–3244. AAAI Press, 2013.

[261] Paul Taele and Tracy Hammond. Developing sketch recognition and interaction techniques for intelligent surfaceless sketching user interfaces. In *Proceedings of the Companion Publication of the 19th International Conference on Intelligent User Interfaces*, pages 53–56. ACM, 2014.

[262] Paul Taele and Tracy Hammond. Enhancing instruction of written east asian languages with sketch recognition-based intelligent language workbook interfaces. In *The Impact of Pen and Touch Technology on Education*, pages 119–126. Springer, 2015.

[263] Paul Taele, Joshua Peschel, and Tracy Hammond. A sketch interactive approach to computer-assisted biology instruction. In *Proc. IUI 2009 Workshop on Sketch Recognition*, 2009.

[264] Yongqiang Tan, Jianguo Zhang, Yanqing Hua, Guozhen Zhang, and HK Huang. Content-based image retrieval in picture archiving and communication systems. In *Proc. SPIE*, pages 614515–614515, 2006.

[265] Jinshan Tang, Rangaraj M Rangayyan, Jun Xu, Issam El Naqa, and Yongyi Yang. Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. *Information Technology in Biomedicine, IEEE Transactions on*, 13(2):236–251, 2009.

[266] Erik L Thurfjell, K Anders Lernevall, and AA Taube. Benefit of independent double reading in a population-based mammography screening program. *Radiology*, 191(1):241–244, 1994.

[267] Georgia Tourassi, Sophie Voisin, Vincent Paquit, and Elizabeth Krupinski. Investigating the link between radiologists gaze, diagnostic decision, and image content. *Journal of the American Medical Informatics Association*, 20(6):1067–1075, 2013.

[268] Georgia Tourassi, Hong-Jun Yoon, Songhua Xu, Garnetta Morin-Ducote, and Kathy Hudson. Comparative analysis of data collection methods for individualized modeling of radiologists' visual similarity judgments in mammograms. *Academic Radiology*, 20(11):1371 – 1380, 2013.

[269] Georgia D Tourassi, Brian Harrawood, Swatee Singh, Joseph Y Lo, and Carey E Floyd. Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms. *Medical Physics*, 34(1):140–150, 2007.

[270] Georgia D Tourassi, Rene Vargas-Voracek, David M Catarious Jr, and Carey E Floyd Jr. Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information. *Medical Physics*, 30(8):2123–2130, 2003.

[271] J.D. Tubbs. A note on binary template matching. *Pattern Recognition*, 22(4):359 – 365, 1989.

[272] JHM Tulen, P Moleman, HG Van Steenis, and F Boomsma. Characterization of stress reactions to the stroop color word test. *Pharmacology Biochemistry and Behavior*, 32(1):9–15, 1989.

[273] Douglas Tweed and Tutis Vilis. Geometric relations of eye position and velocity vectors during saccades. *Vision Research*, 30(1):111 – 127, 1990.

[274] Nash Unsworth and Randall W Engle. The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, 114(1):104, 2007.

[275] Stephanie Valentine and Martin Field. A shape comparison technique for use in sketch-based tutoring systems. In *Proceedings of the 2011 Intelligent User Interfaces Workshop on Sketch Recognition (Palo Alto, CA, USA, 2011), IUI*, 2011.

[276] Stephanie Valentine, Raniero Lara-Garduno, Julie Linsey, and Tracy Hammond. *The Impact of Pen and Touch Technology on Education*, chapter Mechanix: A Sketch-Based Tutoring System that Automatically Corrects Hand-Sketched Statics Homework, pages 91–103. Springer International Publishing, Cham, 2015.

[277] Stephanie Valentine, Francisco Vides, George Lucchese, David Turner, Honghoe Kim, Wenzhe Li, Julie Linsey, and Tracy Hammond. Mechanix: a sketch-based tutoring and grading system for free-body diagrams. *AI Magazine*, 34(1):55, 2012.

[278] Stephanie Valentine, Francisco Vides, George Lucchese, David Turner, Honghoe Kim, Wenzhe Li, Julie Linsey, and Tracy Hammond. Mechanix: A sketch-based tutoring system for statics courses. In *Proceedings of the 24th Innovative Applications of Artificial Intelligence Conference on Artificial Intelligence (IAAI)*, pages 2253–2260. AAAI, 2012.

[279] HG Van Steenis and JHM Tulen. The effects of physical activities on cardiovascular variability in ambulatory situations [ecg/accelerometry analysis]. In *Engineering in Medicine and Biology Society, 1997. Proceedings of the 19th*

Annual International Conference of the IEEE, volume 1, pages 105–108. IEEE, 1997.

[280] S. Voisin, Hong-Jun Yoon, G. Tourassi, G. Morin-Ducote, and K. Hudson. Personalized modeling of human gaze: Exploratory investigation on mammogram readings. In *Biomedical Sciences and Engineering Conference (BSEC), 2013*, pages 1–4, May 2013.

[281] Sophie Voisin, Frank Pinto, Garnetta Morin-Ducote, Kathleen B Hudson, and Georgia D Tourassi. Predicting diagnostic error in radiology via eye-tracking and image analytics: Preliminary investigation in mammography. *Medical physics*, 40(10):101906, 2013.

[282] Sophie Voisin, Frank Pinto, Songhua Xu, Garnetta Morin-Ducote, Kathy Hudson, and Georgia D Tourassi. Investigating the association of eye gaze pattern and diagnostic error in mammography. In *Proc. SPIE*, pages 867302–867302, 2013.

[283] Tanja GM Vrijkotte, Lorenz JP Van Doornen, and Eco JC De Geus. Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability. *Hypertension*, 35(4):880–886, 2000.

[284] Vesna Vuksanović and Vera Gal. Heart rate variability in mental stress aloud. *Medical Engineering & Physics*, 29(3):344–349, 2007.

[285] Christopher D Wickens. Processing resources and attention. *Multiple-Task Performance*, pages 3–34, 1991.

[286] Christopher D Wickens. Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3(2):159–177, 2002.

[287] Glenn F Wilson. An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *The International Journal of Aviation Psychology*, 12(1):3–18, 2002.

[288] Svante Wold, Henrik Antti, Fredrik Lindgren, and Jerker Öhman. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 44(1):175–185, 1998.

[289] Aaron Wolin, Brian Eoff, and Tracy Hammond. Shortstraw: A simple and effective corner finder for polylines. In *Proceedings of Sketch-Based Interfaces and Modeling (SBIM)*, 2008.

[290] Aaron Wolin, Brian Eoff, and Tracy Hammond. Search your mobile sketch: Improving the ratio of interaction to information on mobile devices. In *Proc. IUI 2009 Workshop on Sketch Recognition*, 2009.

[291] Aaron Wolin, Martin Field, and Tracy Hammond. Combining corners from multiple segmenters. In *Proceedings of the Eighth Eurographics Symposium on Sketch-Based Interfaces and Modeling*, pages 117–124. ACM, 2011.

[292] Aaron Wolin, Brandon Paulson, and Tracy Hammond. Eliminating false positives during corner finding by merging similar segments. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, pages 1836–1837. AAAI Press, 2008.

[293] Aaron Wolin, Brandon Paulson, and Tracy Hammond. Sort, merge, repeat: an algorithm for effectively finding corners in hand-sketched strokes. In *Proceedings of the 6th Eurographics Symposium on Sketch-Based Interfaces and Modeling*, pages 93–99. ACM, 2009.

[294] Aaron David Wolin. *Segmenting hand-drawn strokes*. PhD thesis, Texas A&M University, 2010.

[295] Dongrui Wu, Christopher G Courtney, Brent J Lance, Shrikanth S Narayanan, Michael E Dawson, Kelvin S Oie, and Thomas D Parsons. Optimal arousal identification and classification for affective computing using physiological signals: virtual reality stroop task. *Affective Computing, IEEE Transactions on*, 1(2):109–118, 2010.

[296] Songhua Xu and Georgia Tourassi. A novel local learning based approach with application to breast cancer diagnosis. In *Proc. SPIE*, pages 83151Y–83151Y, 2012.

[297] Bonnie C Yankaskas, Michael J Schell, Richard E Bird, and David A Desrochers. Reassessment of breast cancers missed during routine screening mammography: a community-based study. *American Journal of Roentgenology*, 177(3):535–541, 2001.

[298] Alfred L. Yarbus. *Eye Movements and Vision*, chapter Eye Movements During Perception of Complex Objects, pages 171–211. Springer US, Boston, MA, 1967.

[299] Lexiang Ye and Eamonn Keogh. Time series shapelets: A new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 947–956, New York, NY, USA, 2009. ACM.

[300] Lexiang Ye and Eamonn Keogh. Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. *Data Mining and Knowledge Discovery*, 22(1):149–182, 2010.

[301] Hong-Jun Yoon, Tandy R Carmichael, and Georgia Tourassi. Gaze as a biometric. In *Proc. SPIE*, pages 903707–903707, 2014.

[302] Hong-Jun Yoon, Tandy R. Carmichael, and Georgia Tourassi. Temporal stability of visual search-driven biometrics. In *Proc.*, volume 9416, pages 94160U–94160U–7, 2015.

[303] Robert E Yoss, Norma J Moyer, and Robert W Hollenhorst. Pupil size and spontaneous pupillary waves associated with alertness, drowsiness, and sleep. *Neurology*, 20(6):545–545, 1970.

[304] Songyang Yu and Ling Guan. A cad system for the automatic detection of clustered microcalcifications in digitized mammogram films. *Medical Imaging, IEEE Transactions on*, 19(2):115–126, 2000.

[305] Robert Zeleznik and Timothy Miller. Fluid inking: Augmenting the medium of free-form inking with gestures. In *Proceedings of Graphics Interface 2006*, GI '06, pages 155–162, Toronto, Ont., Canada, Canada, 2006. Canadian Information Processing Society.

APPENDIX A

REMOVAL OF SUBJECT-DEPENDENT AND ACTIVITY-DEPENDENT VARIATION IN PHYSIOLOGICAL MEASURES OF STRESS*

A.1 Abstract

The ability to monitor stress levels in daily life can provide valuable information to patients and their caretakers, help identify potential stressors, determine appropriate interventions, and monitor their effectiveness. Wearable sensor technology makes it now possible to measure non-invasively a number of physiological correlates of stress, from skin conductance to heart rate variability. These measures, however, show large individual differences and are also correlated with the physical activity of the subject. In this paper, we propose two multivariate signal processing techniques to reduce the effect of both forms of interference. The first method is an unsupervised technique that removes any systematic variation that is orthogonal to the dependent variable, in this case physiological stress. In contrast, the second method is a supervised technique that first projects the data into a subspace that emphasizes these systematic variations, and then removes them from the data. The two methods were validated on an experimental dataset containing physiological recordings from multiple subjects performing physical and/or mental activities. When compared to z-score normalization, the standard method for removing individual differences, our methods can reduce stress prediction errors by as much as 50%.

*Reprinted with permission from "Removal of subject-dependent and activity-dependent variation in physiological measures of stress," by F. Alamudun; J. Choi; R. Gutierrez-Osuna; H. Khan; B. Ahmed, 2012. *6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth),* pp. 115-122, Copyright 2012 by IEEE.

A.2 Introduction

The increased occurrence of stress-related illnesses in the United States has resulted in an increased demand for diagnosis, treatment, and management from hospitals and healthcare practitioners [18]. Unfortunately, this trend is unsustainable given traditional healthcare models. For this reason, there has been a push for proactive healthcare technologies that reduce the burden on the healthcare system [203]. This is particularly important in stress management due to its nature: stress monitoring requires extensive patient observation (in his/her natural environment) in order to identify stress triggers and effective interventions. Towards this end, recent advances in wearable sensors allow capturing of various bio-signals non-invasively with minimal impact on patients routines. Such bio-sensors provide health-management options, extending observation and diagnosis beyond the confines of clinical facilities.

A variety of physiological signals have been shown to correlate with stress levels [42, 33], including electrodermal activity (EDA), heart rate (HR), various indexes of heart rate variability (HRV), blood volume pressure (BVP), pupil dilation, muscle tension, and respiration. However, much of this work has been performed under controlled laboratory settings, and only a few studies have investigated ambulatory stress monitoring [109, 251, 132]. When monitoring in ambulatory domains, subtle physiological responses to psychological stress can be easily masked by various interferences, from changes in posture and physical activity (e.g., walking) to environmental factors (e.g., temperature). An added challenge: physiological baselines and physiological responses to stressors are highly individual. As a result, mapping physiological signals into stress indexes may require substantial calibration data from each patient.

To address these issues, the work presented here explores two multivariate tech-

niques to reduce the effects that physical activity and individual differences have on physiological stress responses. The first technique is based on an orthogonal signal correction (OSC) algorithm originally developed by Wold et al. [288] to remove systematic noise in near-infrared (NIR) spectroscopy. The second method is based on the classical Fishers linear discriminant analysis (LDA). In this method, we use LDA to identify the main directions of variance for each interference (individual difference or physical activity), and then subtract them from the original data matrix through least squares.

The paper is organized as follows. First we provide a brief background review of bio-signals and prior work on wearable platforms. Next, we describe the two signal-denoising methods. We then provide a brief overview our experimental protocol for eliciting psycho-physiological responses under physical and mental stressors. Finally, results from the experimental comparison of the two denoising methods are provided, followed by a discussion of results and directions for future work.

A.3 Background

A.3.1 Stress and the Human Body

Stress describes the physiological changes that occur in response to a category of perceived physical or psychological threats. Under normal conditions, stress helps to keep the body alert and composed to avert any threats. However, if the frequency or the duration of the stressor is excessive, brief or prolonged, stress responses can lead to long-term health complications for an individual [42]. Stress has been linked to numerous cardiovascular diseases, immunosuppression and hypertension and to psychological disorders such as anxiety, difficulty assimilating new information and increased dependence on and abuse of alcohol and drugs [113, 173, 272].

There are two elements involved in the human reaction to stress, the hypothalamic-

pituitary-adrenal (HPA) axis, and the sympathetic nervous system (SNS), which along with the parasympathetic nervous system (PNS) form branches of the autonomic nervous system (ANS) [42]. The ANS is that part of the nervous system that controls involuntary functions. The parasympathetic and sympathetic branches counteract each other and serve to balance each other at the same time: the sympathetic branch controls the activation of stress or flight or fight response, whereas the parasympathetic branch promotes relaxation and energy conservation.

A.3.2 Physiological Stress Response

The organs of the body (including cardiac and respiratory organs) are connected to both autonomic branches with the exception of the skin. Skins blood vessels and eccrine sweat glands are exclusively enervated by the sympathetic nervous system [33]. For this reason, changes in the active and passive electrical properties of the skin are an ideal measure of sympathetic activation and therefore of stress. These changes are commonly referred to as electrodermal activity (EDA). There exists a large body of research on EDA, including its use as a measure of stress [218, 244, 126, 221]. However, relying solely on measures of skin conductance as a sole marker for sympathetic nervous system activation is insufficient for stress monitoring. For example, skin conductance also increases in response to physical exertion due to increased eccrine sweat gland activation [239]. Other physiological measures, such as cardiac activation and respiration rate, although influenced by both autonomic branches, provide complementary information that may be useful in discriminating the stress response in an organism [48, 49, 284, 219, 283].

A.3.3 Factors Affecting Physiological Stress Response

With a few exceptions[221, 236, 250] most previous research on stress detection has focused on controlled laboratory environment or semi-controlled ambulatory set-

tings, where subjects are constrained to a sedentary posture. While these studies provide the fundamentals for understanding the psycho-physiology of stress, they do not account for factors that are encountered in real-world scenarios where subjects seldom maintain the same posture or have restricted movement. In ambulatory settings, an individual adapts internally in response to changes in level of physical exertion and posture. According to Olufsen et al. [199], the cardiovascular responses to postural change from sitting to standing and from standing to varied levels of movement involve interactions between the autonomic nervous system, which regulates heart rate, perspiration and pupil dilation, and cerebral autoregulation.

Van Steenis et al. quantified posture-related changes in heart rate [279]. In this work, the authors reported a significant increase in heart rate as a subject transitions from supine to sitting, from sitting to standing, and from standing to walking. In their work on long-term monitoring using EDA, Kappeler-Setz et al. [126] measured the effect of movement in a single limb on skin conductance response on fingers and feet, and concluded that these effects were minor. However, in a study on the effect of full body movement on EDA, Schumm et al. [242] concluded that the faster a person is walking the more the peak distribution of skin conductance response approaches a uniform distribution. They also concluded that at walking speeds in excess of 6 km/h (3.72 mph) the probability of detecting EDA in response to specific events is significantly decreased.

A number of research studies [48, 49, 295] have concluded that these differences make it difficult to translate information from one subject to inferences about the state of another subject. As an example, Wu et al. [295] reported a 61.8% drop in predictive accuracy, from 96.5% for subject-dependent models down to 36.9% for subject-independent models, when classifying arousal levels using a set of biosignals (skin conductance level, respiration, ECG, and EEG). In summary, changes

in physical activity and individual differences across subjects can mask the effect of psychological stress on physiological signals. To address these issues, the work presented here describes two multivariate methods that may be used to subtract these interferences from raw physiological signals. If successful, these methods may pave the way towards the development of subject-independent stress monitoring in ambulatory settings.

A.4 Reducing Individual Differences and Effects from Physical Activity

In real-world scenarios two sources of variance can contaminate the physiological signals during stress monitoring: individual differences in physiological baseline and physiological stress response, and physiological responses due to physical activity. The methods proposed in this work assume that both influences can be treated as systematic noise sources that are independent from the observation of interest. Specifically, we propose two multivariate filtering strategies: orthogonal signal correction (OSC) and linear discriminant correction (LDC). The first method (OSC) assumes that systematic noise components are orthogonal to the variation of interest (stress response) and applies a filter to remove all components orthogonal to the latter. The second method (LDC) uses Fishers linear discriminant analysis (LDA) to model each individual systematic noise component iteratively and then subtracts it from the raw physiological response. The result of both methods is a physiological signal where psychological stress is more salient.

A.4.1 Orthogonal Signal Correction

The concept of orthogonal signal correction (OSC) was originally introduced by Wold et al. [288] as a pre-processing step for removal of systematic noise such as baseline variation and multiplicative scatter effects in near-infrared (NIR) spectra. The method was later generalized by Fearn [71] to other NIR applications. In our

work, OSC is applied to remove sources of systematic variation from physiological response data that are uncorrelated with (orthogonal to) the applied stress stimuli. Thus, by treating variation introduced by sources other than stress stimuli as structured noise, an OSC filter can be used as a pre-processing step to remove such noise.

This process is accomplished by constraining the removal of components from the physiological response data to only those components that are orthogonal to the applied stress stimuli. For this purpose, we decompose the physiological response data into correlated and uncorrelated factors:

$$X = X_{\overline{C}} + X_C \tag{A.1}$$

where X is a matrix of physiological responses,
Y is a vector of target variables (applied stress stimuli),
X_C is the response correlated with the applied stress stimuli, and
$X_{\overline{C}}$ is the response uncorrelated with the applied stress stimuli ($X_{\overline{C}} \perp Y$).

The OSC algorithm expresses the data matrix X in bilinear form:

$$X = tp^T + E \tag{A.2}$$

where X is the $(N \times K)$ matrix of unfiltered data,
E is the $(N \times K)$ matrix of noise (in our case the "filtered" data),
t is a $(N \times L)$ score matrix, and
p is a $(K \times L)$ matrix of loadings.

The number of samples and variables of the "training set" (calibration set) are N and K respectively, and L is the number of components (latent variables). The objective of OSC is to find t and p subject to the orthogonality constraint:

$$t \perp Y \tag{A.3}$$

Wold et al. [288] use an iterative procedure to calculate t:

$$t = (1 - Y(Y^T Y)^{-1} Y^T)t \tag{A.4}$$

which is orthogonal to Y since:

$$Y^T t = Y^T (1 - Y(Y^T Y)^{-1} Y^T)t \tag{A.5}$$

$$(Y^T - Y^T Y(Y^T Y)^{-1} Y^T)t = 0 \tag{A.6}$$

where t is initialized using the first principal component of X.

After each iteration, the convergence is checked by comparing the difference between the newly predicted t and the previous t. The target value for t is obtained when this difference converges to a value below a predetermined threshold. From here, the loading vector p is calculated as:

$$p = (X^T t(t^T t)^{-1} \tag{A.7}$$

After convergence, the uncorrelated (i.e., noise) and correlated (i.e., signal) compo-

nents can be obtained by Equation A.1 and A.2 as:

$$X_{\overline{C}} = tp^T \tag{A.8}$$

$$X_C = X - X_{\overline{C}} \tag{A.9}$$

A.4.2 Linear Discriminant Correction

Fishers linear discriminant analysis (LDA) is a transformation that seeks to determine a low-dimensional projection where the separation between two or more classes is maximized [65]. Thus, LDA can be used to find a projection where subject-to-subject differences (or effects from physical activity) are maximized. Once this low-dimensional projection is found, it can be subtracted in a multivariate fashion from the full data matrix by means of least-squares regression. As a result, the deflated data matrix will not contain any of the variability in the LDA projection. This process is applied iteratively to each noise source. We refer to the resulting algorithm as the *linear discriminant correction* (LDC) method.

In what follows, we describe the process of removing subject-to-subject differences; the process for removing physical activity is identical. As before, we assume that the physiological response matrix X (for multiple subjects) can be decomposed as:

$$X = X_{\overline{C}} + X_C \tag{A.10}$$

where X_C is the filtered response correlated with the applied stimuli and
$X_{\overline{C}}$ is the uncorrelated response (e.g., individual differences).

To estimate $X_{\overline{C}}$, we apply LDA to matrix X using the subjects identity $\omega_{ID} =$

$1, 2, m$ as class labels (m being the number of subjects):

$$(X, \omega_{ID}) \xrightarrow{\text{LDA}} t_{\omega_{ID}}, p_{\omega_{ID}} \tag{A.11}$$

where $p_{\omega_{ID}}$ is a matrix of loadings (or eigenvectors)
and $t_{\omega_{ID}}$ denotes the score matrix (projection of the data onto the eigenvectors):

$$t_{\omega_{ID}} = X p_{\omega_{ID}} \tag{A.12}$$

The score matrix is a subspace in which individual differences across subjects are maximized. Next, we predict the full data matrix X from $t_{\omega_{ID}}$ as:

$$X = \beta_{\omega_{ID}} t_{\omega_{ID}} \tag{A.13}$$

where $\beta_{\omega_{ID}}$ is a vector of regression coefficients, which can be estimated by:

$$\beta_{\omega_{ID}} = \underset{\beta}{argmin} \, ||X - \beta_{\omega_{ID}} t_{\omega_{ID}}|| \tag{A.14}$$

$$\beta_{\omega_{ID}} = (t_{\omega_{ID}}^T t_{\omega_{ID}})^{-1} t_{\omega_{ID}}^T X \tag{A.15}$$

Hence, $X_{\overline{C}}$ becomes:

$$\beta_{\omega_{ID}} = \beta_{\omega_{ID}} X t_{\omega_{ID}}) \tag{A.16}$$

By subtracting $X_{\overline{C}}$ from X, we then obtain a matrix X_C where individual differences across subjects have been minimized:

$$X_C = X - X_{\overline{C}} \tag{A.17}$$

The same process is repeated to remove the effect of physical activity: we perform LDA to project X_C onto a subspace in which physical activity differences are maximized, and then subtract this information as we did in Equations A.11 – A.17.

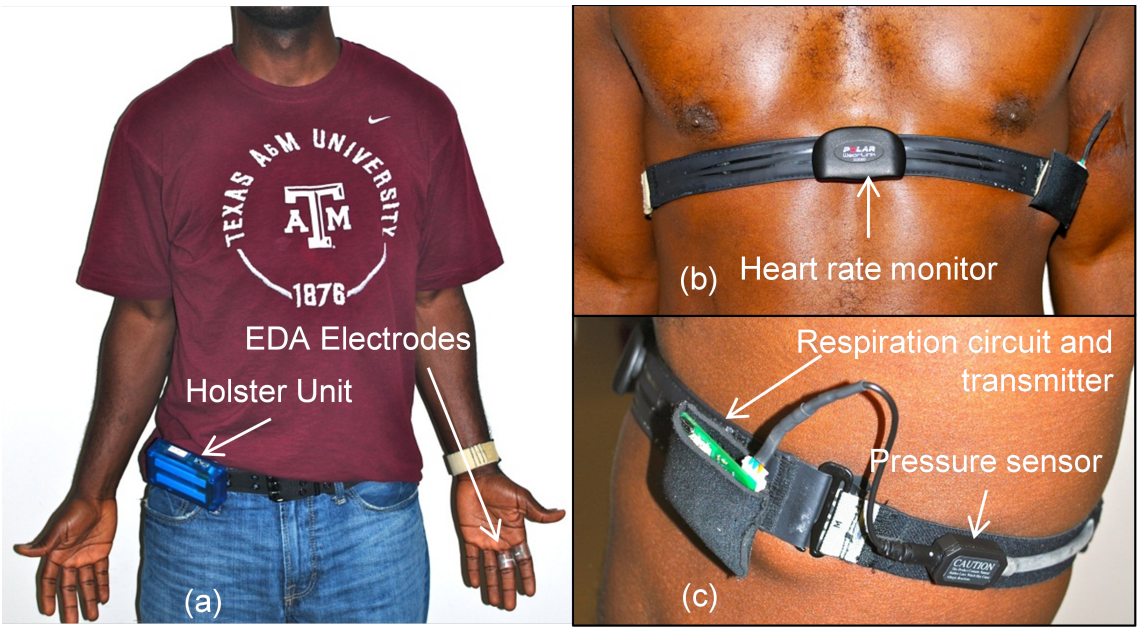


Figure A.1: Wearable sensor prototype. (a) Subject wearing complete system with visible holster unit, two electrodes placed on the proximal phalanges of middle and index finger, the wireless EDA node is placed on the wrist band. (b) The HRM is located on the center of the chest. (c) Respiration sensor and transmitter is located on the left side of the chest (from [3]) et al. [3]).

A.5 Materials and Methods

We evaluated the proposed noise-cancellation methods on experimental data from a pool of participants. In the experiment, we recorded the participants reaction to three mental tasks, each of which elicited different arousal levels, while varying the participants physical activity.

A.5.1 Wearable Sensor System

For these experiments, we used a wearable sensor system that has been developed by our group over the course of the past three years [48, 49, 50]. The system consists of a heart rate monitor, a respiration sensor and an EDA sensor. Heart rate was measured with a Polar WearLink+ heart-rate-monitor (HRM) (Polar Electro Inc.), whereas respiratory activity was measured with a pressure-based respiration sensor (SA9311M, Thought Technology Ltd) integrated in the HRM chest strap. Finally, we measured EDA in a constant-voltage configuration using two electrodes on the proximal phalanges of the index and the middle finger of the non-dominant hand. Small AgCl electrodes (E243; In Vivo Metric Systems Corp.) were used for this purpose.

Sensor signals are wirelessly transmitted to a holster unit containing an embedded Linux microcontroller (Marvell PXA270 400 MHz, 64 MB RAM; Gumstix, Inc.), a heart rate receiver module (RMCM01; Polar Electro Inc.) and a wireless transceiver to communicate with the respiration and EDA sensor. The sensor hub is also responsible for power management of the holster unit, a 3000 mAh Li-Po battery, which allows for data to be continuously collected for over thirteen hours. Figure A.1 shows the sensor configuration and placement.

A.5.2 Experimental Setup

The experimental protocol consisted of four sessions (sitting, standing, slow walking, and fast walking), each representing a unique posture or physical activity level; see Table A.1. Fourteen volunteers (age range: 18 - 35) were asked to participate in the experiment after giving informed consent. Subjects reported that they were in good health; none reported excessive drinking or smoking habits. They were requested not to undertake unusual activities such as heavy training or abnormal drinking a

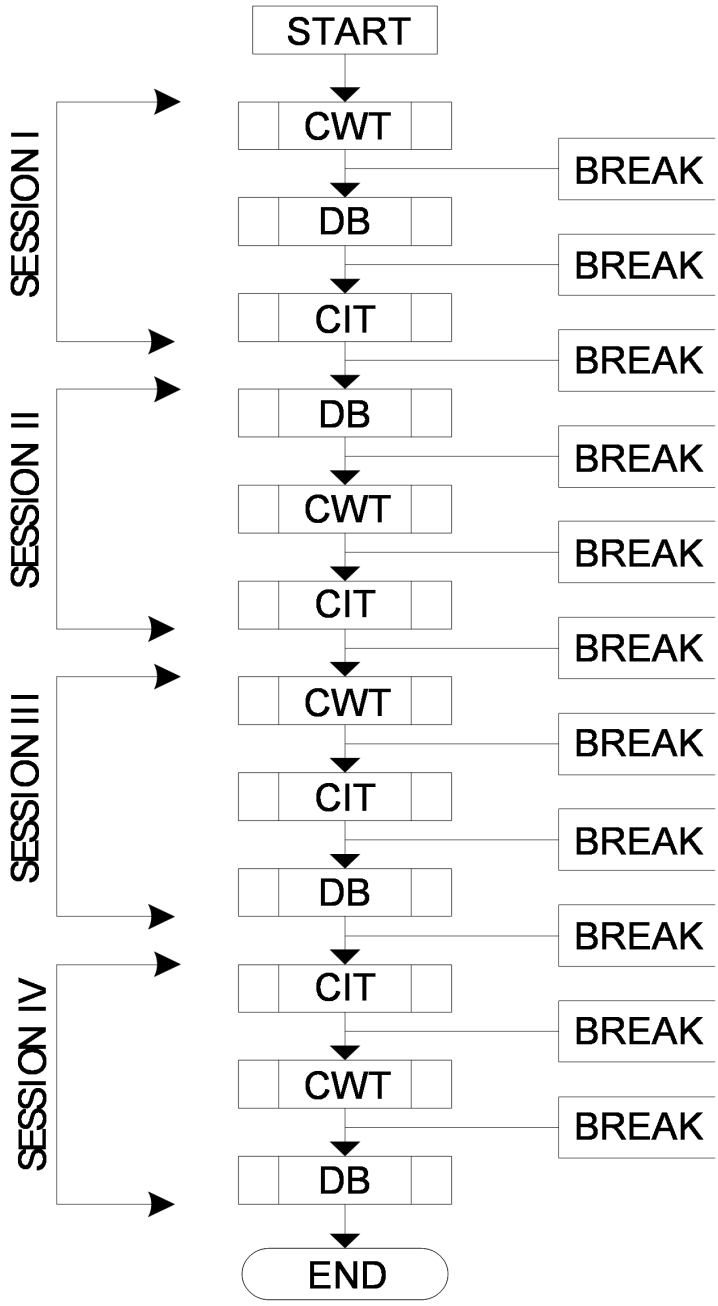


Figure A.2: Experimental protocol The CWT, CIT and DB tasks lasted 5, 3 and 2 minutes respectively with a 2 minute break between tasks. Each task was repeated during all four sessions.

Table A.1: Summary of experimental protocol.

Session	Session Description
I - Sitting	Subjects were required to remain seated in an upright position on an immobilized chair. The chair was adjusted to the size of each subject prior to start of the experiment.
II - Standing	Subjects were required to remain standing in an upright position.
III - Slow walking	Subjects were placed on a treadmill and asked to maintain a constant slow walking pace of 1.24 mph.
IV - Fast walking	Subjects were placed on a treadmill and asked to maintain a constant fast walking pace of 2.17 mph.

day prior to the experimental sessions. Subjects were also asked to avoid caffeinated products 6 hours prior to the experimental sessions. The experimental protocol and procedures in this study were approved by the Texas A&M University Institutional Review Board.

From the fourteen subjects who volunteered for the study, 12 were male and 2 were female. Data from two of the subjects (one male, one female) was excluded due to sensor noise and wireless connectivity issues during data collection. The second female subject was also excluded to maintain homogeneity in the dataset (i.e., gender).

An overview of the experimental protocol is shown in Figure A.2. For each session, subjects were asked to perform three tasks: one eliciting high stress, one eliciting low stress, and a controlled relaxation task; presentation of each task was randomized for each subject. After each task, subjects had a 2 minute break period for recovery. For the high-stress task, subjects were subjected to 5 minutes of a mobile version of the Stroop color word conflict test (CWT). In the conventional CWT, the participant is shown one of four words (Red, Green, Blue, Yellow) displayed in a different ink color, and asked to respond based on the ink color; e.g., in the example shown in Figure A.3(b) the correct answer is Green.

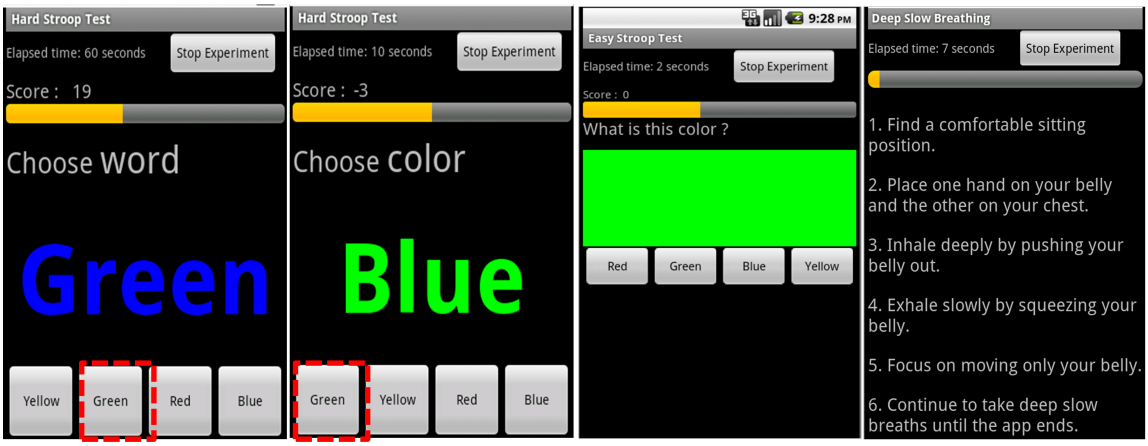


Figure A.3: Android smartphone platform based tasks. (a) CWT task word name prompt. (b) CWT task ink color prompt. (c) CIT task. (d) DB task.

We introduced two variations to make the CWT more challenging and minimize learning effects [123]. First, rather than always asking participants to respond to the ink color, 50% of the times they had to respond to the word; see Figure A.3(a). This forces participants to switch strategies and makes the test significantly more challenging, Second, the location of the answer buttons at the bottom of the screen is randomized with each word presentation, and a loud bell is played every time the participant choose an incorrect answer. This CWT task was administered via an Android^{TM} mobile device (Figure A.3). The second task was designed to elicit a lower stress reaction in comparison to the first task. Subjects were subjected to three minutes of a color identification test (CIT). During this task, subjects were asked to confirm a displayed color; see Figure A.3(c). This task was also presented on an Android^{TM} mobile device. For the third task, participants were asked to perform a deep breathing (DB) relaxation exercise for two minutes; instructions were provided as shown in Fig 3(d).

Upon completion of each task, subjects were asked to provide a self-reported evaluation of arousal level. All other factors such as room temperature, humidity and

sunlight/light intensity were kept constant for all subjects throughout the experimental procedure.

Using data collected from 11 subjects who participated in the experiment, we extracted a total of 7 features including 2 features from EDA and 5 features from HRV [295]. Respiratory features were not included in the study since they provide a misleadingly high discrimination between deep breathing and CWT. A summary of each feature is provided in Table A.2. Each feature was calculated using a 60s moving window with a 10s shift. All features were normalized to z-scores, the standard method for handling individual differences in skin conductivity [24].

Table A.2: Features extracted from psycho-physiological sensors.

Sensor	**Feature**	**Description**
EDA	μ_{SCL}	Average skin conductance level
	σ_{SCL}	Standard deviation in skin conductance level
Heart Rate	LF_{HRV}	Low frequency power in HRV (0.04 - 0.15 Hz)
	HF_{HRV}	High frequency power in HRV (0.15 - 0.5 Hz)
	$LF:HF_{HRV}$	Ratio of LF to HF power content in HRV
	$AVNN$	Average of R-R intervals
	$SDNN$	Standard deviation of successive R-R intervals

A.6 Results

As a first step in analyzing the physiological responses, we compared the average skin conductance level (SCL) and the average R-R intervals (AVNN) for each of the performed tasks. Results are shown in Figure A.4. As anticipated, (1) the CWT invoked an increase in heart rate (a reduction in R-R interval) and a significant increase in average skin conductance, (2) the CIT task invoked a lower response in comparison with the CWT task, and (3) the DB relaxation task invoked a reduction in heart

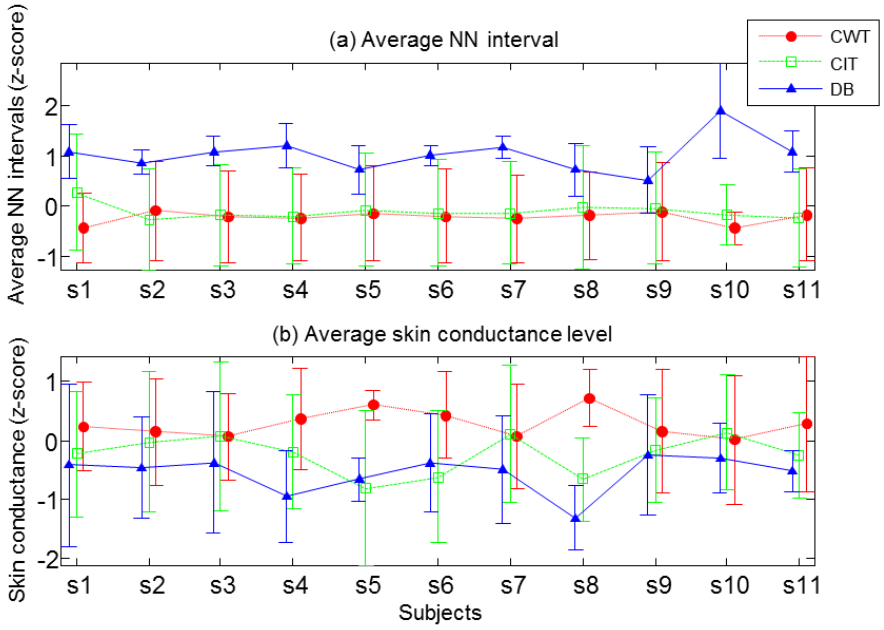


Figure A.4: Android smartphone platform based tasks. (a) CWT task word name prompt. (b) CWT task ink color prompt. (c) CIT task. (d) DB task.

rate (an increase in R-R interval) and a reduction in average skin conductance. These results provide evidence for the validity of our experimental protocol.

To analyze the effect of physical activity on a subjects physiological response to each task, we compared the skin conductance level (SCL) and R-R intervals (AVNN) for each of the performed tasks. Results shown in Figure A.5 indicate that there was a negligible difference in average heart rate between the sitting and standing postures. As expected, there was a significant increase in average heart rate and average skin conductance level when the subject was mobile (slow walking or fast walking).

Finally, we analyzed the effectiveness of the proposed noise-cancellation methods in improving the stress detection across subjects and across physical activity levels. For this purpose, we divided our analysis into three cases: subject-independent, activity-independent, and subject-and-activity independent. On each case, we set up

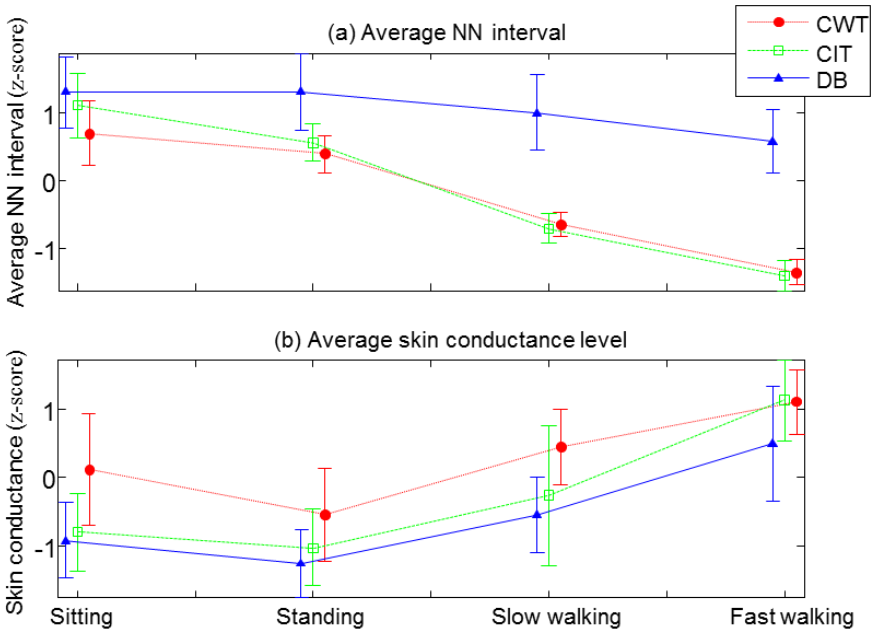


Figure A.5: Comparison of (a) average NN interval (AVNN) and (b) average skin conductance level (SCL) across all subjects.

a binary classification problem, with the CWT condition as the stress class and the DB condition as the no-stress class. To generate a balanced dataset (CWT lasted for 5 min whereas DB lasted for 2 min), we randomly selected (without replacement) an equal number of analysis windows from each session. This random sampling process was repeated 50 times; classification rate reported here is the average across the 50 runs.

In the subject-independent case we studied whether the stress level of one subject (CWT: stress; DB: no stress) could be predicted from the response of another subject to the same set of tasks, given that both subjects maintained the same type of physical activity. We used a leave-one-subject-out cross validation approach whereby data from each subject was used for testing a model trained on the remaining sub-

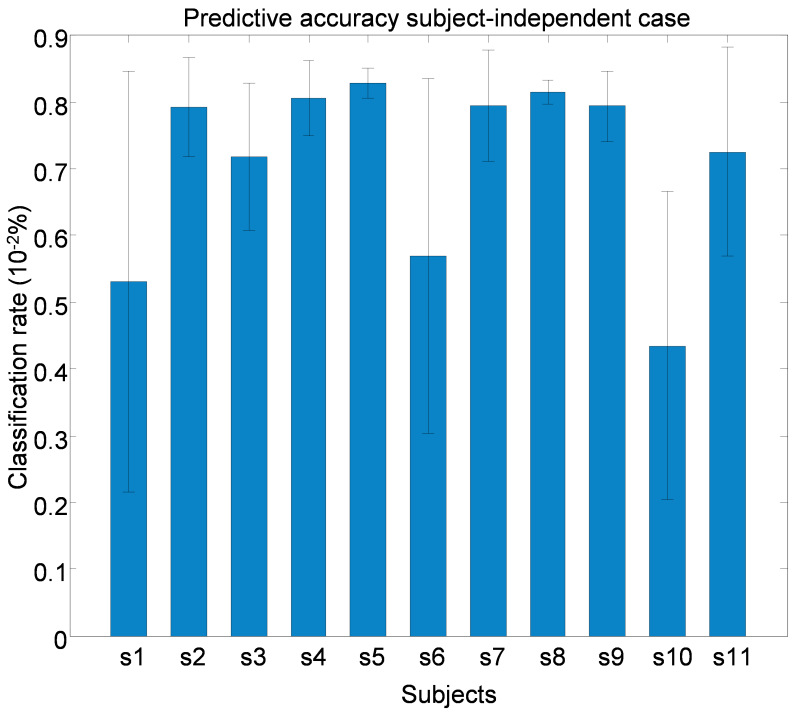


Figure A.6: Average classification rate for subject-independent case ($\mu = 0.67$, $\sigma = 0.19$).

jects. For each subject, four quadratic classifiers were trained to discriminate between the CWT and DB tasks, one classifier per session (sitting, standing, slow walk, fast walk), for a total of 44 models (11 subjects 4 activities). Figure A.6 shows the average prediction results for each subject, averaged over the four sessions. These results reveal significant individual differences, with classification performance ranging from 35% ($s10$) to 82% ($s5$). In the self-assessment report, $s10$ indicated that he found the DB task highly stressful during the first session and slightly stressful during the third and fourth sessions. In contrast, subject s5 indicated that he found the CWT task highly stressful and the DB task very relaxing. Thus, difference in classification performance across subjects may be explained (in part) by the fact that subjects can have a radically different experience when performing the same task.

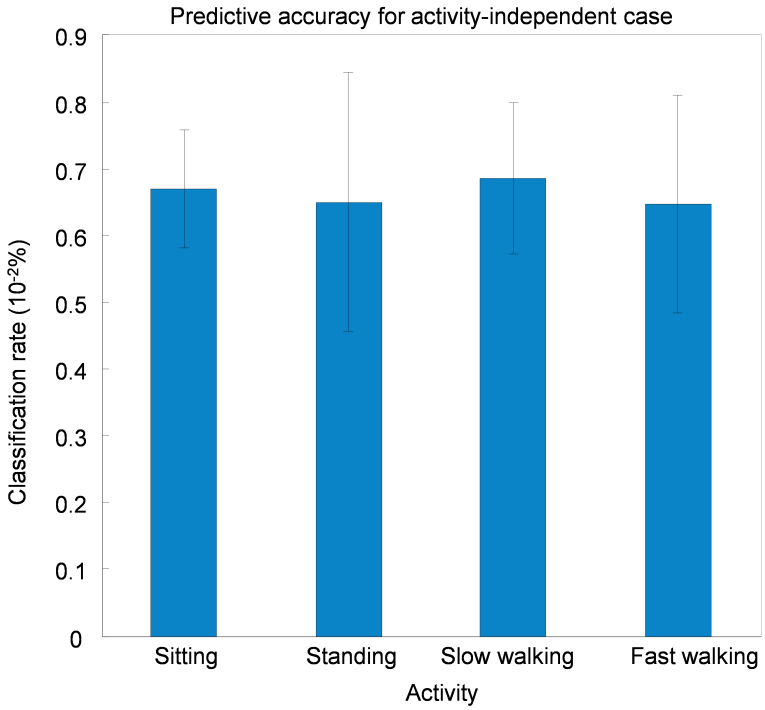


Figure A.7: Average classification rate for activity-independent case ($\mu = 0.66$, $\sigma = 0.14$).

In the activity-independent case we studied whether the stress level of one subject to a set of tasks (CWT: stress; DB: no stress) could be predicted from his/her prior responses to the same set of tasks under different levels of physical activity. We used a within-subject leave-one-session-out cross validation approach, where data from one session was used for testing while data from the remaining sessions was used for training. Thus, a total of 44 models (11 subjects 4 activities) were also trained. Classification results were averaged across the eleven subjects, and are summarized in Figure A.7. As measured by the average classification rate, individual differences and physical activity have comparable effects.

In the subject-and-activity independent case, we studied whether the stress level of one subject (CWT: stress; DB: no stress) could be predicted from the response of

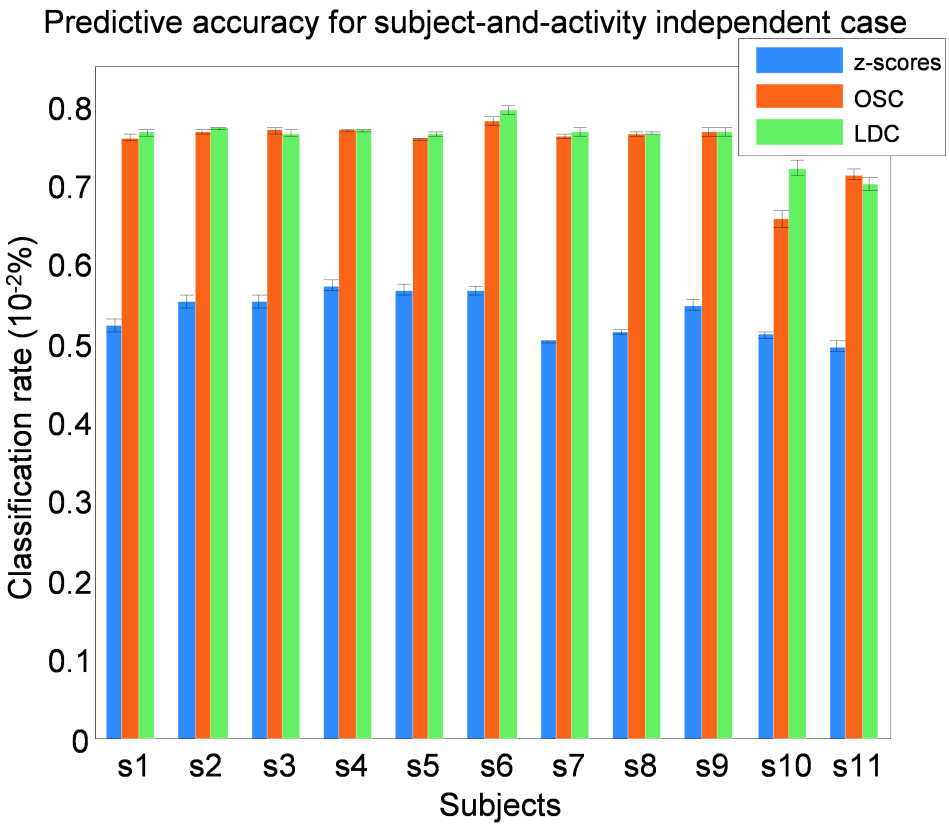


Figure A.8: Average classification rates for subject-and-activity independent case.

another subject, regardless of the physical activity levels. For cross validation, data from each subject (all activity levels) was used for testing a model trained on the remaining subjects (all activity levels). This resulted in 11 classifiers, one per subject. Results are shown in Figure A.8 for classification performance on z-scores vs. those obtained following application of the two noise-cancellation methods.

From Figure A.8, we observe that the average classification results prior to noise-cancellation ($\mu = 53.63, \sigma = 2.9$) are significantly lower than those in the previous two cases, which illustrates the compounding effects of individual differences and physical activity on mental stress detection. Application of the OSC and LDC noise-cancellation methods results in a 48% reduction in error rate (from 46.54% for z-scores to 23.73% on OSC/LDC). OSC was optimized using 2 components, a tolerance value of 99.99, and 100 iterations. The LDC method was implemented using the first 2 eigenvectors for posture variation and the first 3 eigenvectors for subject variation.

To visualize the effect of the noise correction method, we compared the structure of the physiological response for all subjects before and after LDC correction using principal component analysis. From Figure A.9, we observe the distribution of the stress class (CWT) and the no-stress class (DB) using the first two principal components. The application of the noise correction method results in an increased distance between the mean of the two classes.

A.7 Discussion and Conclusions

Differences in physiology across subjects and changes in physical activity can overshadow the subtler physiological responses to mental stressors. In this work, we have presented two pre-processing algorithms that may be used to ameliorate the effect of these two types of interferences, making it easier to detect the effects of stressors. The first method, known as orthogonal signal correction, was originally de-

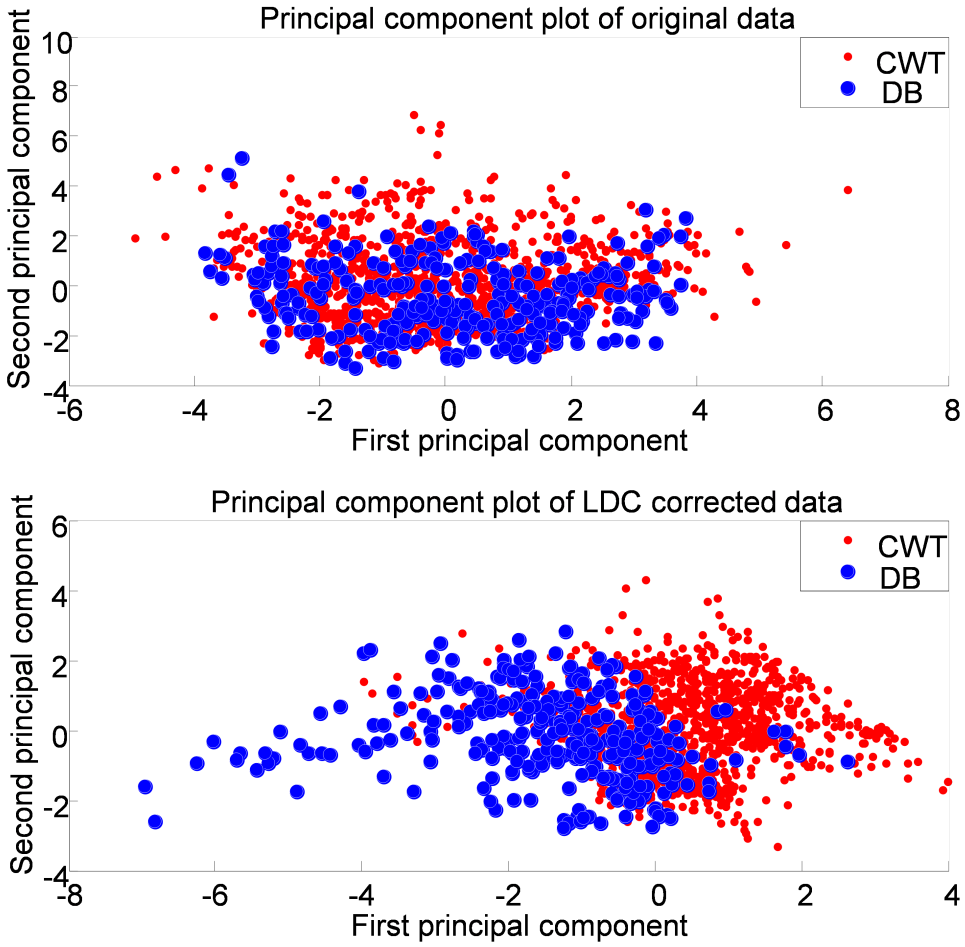


Figure A.9: Principal component analysis of task response (a) before correction and (b) after LDC noise correction.

veloped in the field of chemometrics. OSC attempts to remove any source of variance that is orthogonal to the dependent variable (i.e. stress levels). OSC can be thought of as an unsupervised technique since the specific noise sources need not be identified. In contrast, the second method operates by modeling each unique noise source and then removing it from the data matrix. This approach, which we have termed linear discriminant correction, is based on Fishers linear discriminant analysis.

We validated both methods on experimental data from a number of participant performing three distinct tasks mental tasks (color word test, color identification test, and deep breathing, each of them under four different levels of physical activity (sitting, standing, slow walking, fast walking). For each participant, we computed seven measures related to heart rate variability, electrodermal activity, and respiratory rate, and used the resulting feature vector as independent variables to predict the stress levels (low vs. high stress). In the presence of both sources of variance (individual differences and physical activity) classification performance on z-scores was 53.5%, slightly above chance levels. Following application of noise-cancellation methods, classification performance raised to 75.6% and 76.3%, for OSC and LDC, respectively.

These results indicate that either method can bring noticeable improvements in stress prediction when physiological recordings are affected by changes in physical activity and subject-to-subject differences. Among the two methods, OSC only requires information about the dependent variable (stress levels), whereas LDC requires information about the noise sources in order to estimate their directions of variance. This confers each method its own advantages. OSC is a more general method because it can remove any source of variance not related to the dependent variable. On the other hand, as demonstrated by the results in Figure A.6, obtaining ground-truth for stress levels is problematic since subjects can have a radically different experience

when performing the same task (i.e., some of our subjects perceived deep breathing as being highly stressful). In contrast, LDC only requires ground-truth for subject identity, which can be encoded in the instrument, and physical activity, which can be measured with additional sensors (e.g., accelerometers). The experiments reported in this work were performed in a laboratory setting. Work is underway to evaluate the proposed cancellation methods in ambulatory settings where subjects are allowed to carry on with their daily activities. Results from these experiments will provide a stronger validation on the effectiveness of our methods for real-world stress monitoring applications.

In this paper, we have focused on individual differences and physical activity. Additional research is required to investigate the effect of additional sources of variance that influence physiological stress response such as age, gender, body composition, circadian rhythms etc. Accounting for these and other sources of variance will also need to be considered for real-world applications.

APPENDIX B

TIME SERIES SHAPELET ANALYSIS

Typical classification techniques used in machine learning are not efficient at handling real-valued ordered time series data, where temporal ordering and trends of data are as useful in providing discriminative information as discrete values. Such changes are difficult and often impossible to capture using well-known methods such as nearest neighbor algorithm, where features are individual data points and temporal characteristics are not preserved. These challenges result in the need for a type of data primitive, which (a) capture temporal changes of observed data, (b) generate temporal attributes effective for establishing sufficient criteria for class membership and thus usable as a feature for classification, and (c) can be utilized during post-analyzed for model characteristics responsible for determining class membership.

Shapelets are a time series data mining primitive able to determine similarity between classes based on small common shapes occurring at any point in a series. Finding a shapelet requires generating a set of candidates, defining a distance measure between a shapelet and each time series, and defining a measure of the discriminatory power of a shapelet[299]. We describe Ye and Keogh's algorithm [299] below.

To generate a subsequence S of length l of a time series T of length m is a contiguous sequence of l points in T , where $l \leq m$. Any time series of length m contains $(m - l) + 1$ distinct subsequences of length l. We denote the set of all subsequences of length l for series T_i to be $W_{i,l}$ and the set of all subsequences of length l for data set T to be:

$W_l = W_{1,l}, \ldots, W_{n,l}$.

The set of all candidate shapelets for data set T is:

$$W = W_{min}, W_{min+1}, \ldots, W_{max},$$

where $min \geq 3$ and $max \leq m$. Note that W is very large, with $O(m^3)$ candidate shapelets. Ye and Koegh [299, 300] have proposed efficient pruning of W to improve the time complexity of the exhaustive search, however we only present the generic shapelet finding algorithm here.

```
ShapeletSelection (T, min, max)
1       best = 0;
2       bestShapelet = ;
3       C = classLabels(T );
4       W = generateCandidates(T, min, max);
5               for l = min to max do
6                       for all subsequence S in Wl do
7                               DS = findDistances(S, Wl);
8                               quality = assessCandidate(S, DS );
9                               if quality > best then
10                                      best = quality;
11                                      bestShapelet = S;
12                              end if
13                      end for
14              end for
15      return bestShapelet;
```

APPENDIX C

MAMMOGRAPHIC CASES FROM THE DIGITAL DATABASE FOR SCREENING MAMMOGRAPHY

The following table gives the reference number of the digital database for screening mammography (DDSM) cases used in our study.

Table C.1: Volume and corresponding case number for malignant cases from DDSM

No.	**Volume**	**Case Number**	**No.**	**Volume**	**Case Number**
1	cancer 01	case3022	26	cancer 02	case0070
2	cancer 01	case0001	27	cancer 02	case0073
3	cancer 01	case0003	28	cancer 02	case0082
4	cancer 01	case0004	29	cancer 02	case0089
5	cancer 01	case0006	30	cancer 02	case3023
6	cancer 01	case0014	31	cancer 05	case0031
7	cancer 01	case0016	32	cancer 05	case0085
8	cancer 01	case0017	33	cancer 05	case0128
9	cancer 01	case3010	34	cancer 05	case0140
10	cancer 01	case3012	35	cancer 05	case0142
11	cancer 01	case3018	36	cancer 05	case0143
12	cancer 01	case3033	37	cancer 05	case0146
13	cancer 01	case3057	38	cancer 05	case0148
14	cancer 01	case3073	39	cancer 05	case0149
15	cancer 02	case0018	40	cancer 05	case0155
16	cancer 02	case0027	41	cancer 05	case0156
17	cancer 02	case0032	42	cancer 05	case0157
18	cancer 02	case0034	43	cancer 05	case0158
19	cancer 02	case0035	44	cancer 05	case0160
20	cancer 02	case0038	45	cancer 05	case0161
21	cancer 02	case0040	46	cancer 05	case0164
22	cancer 02	case0041	47	cancer 05	case0165
23	cancer 02	case0042	48	cancer 05	case0168
24	cancer 02	case0043	49	cancer 05	case0170
25	cancer 02	case0059	50	cancer 05	case0175

Table C.2: Volume and corresponding case number for benign cases from DDSM

No.	**volume**	**case number**
1	benign 01	case0217
2	benign 01	case0240
3	benign 01	case0243
4	benign 01	case0245
5	benign 01	case0248
6	benign 01	case0249
7	benign 01	case3093
8	benign 01	case3098
9	benign 01	case3099
10	benign 01	case3100
11	benign 01	case3113
12	benign 01	case3118
13	benign 01	case3128
14	benign 01	case3132
15	benign 01	case3140
16	benign 04	case0251
17	benign 04	case0252
18	benign 04	case0253
19	benign 04	case0273
20	benign 04	case0274
21	benign 04	case0282
22	benign 04	case0283
23	benign 04	case0303
24	benign 04	case0304
25	benign 04	case0306

Table C.3: Volume and corresponding case number for normal cases from DDSM

No.	**volume**	**case number**
1	normal 09	case3601
2	normal 09	case3602
3	normal 09	case3603
4	normal 09	case3604
5	normal 09	case3606
6	normal 09	case3607
7	normal 09	case3608
8	normal 09	case3609
9	normal 09	case3611
10	normal 09	case3612
11	normal 09	case3613
12	normal 09	case3615
13	normal 09	case3618
14	normal 09	case3619
15	normal 09	case3621
16	normal 10	case3660
17	normal 10	case3661
18	normal 10	case3662
19	normal 10	case3663
20	normal 10	case3664
21	normal 10	case3665
22	normal 10	case3666
23	normal 10	case3667
24	normal 10	case3668
25	normal 10	case3670